Using Eye Tracking for Evaluation of Information Visualisation in Web Search Interfaces

A thesis submitted for the degree of

Doctor of Philosophy

Hilal Ali Al Maqbali B.E. M.Sc,

School of Computer Science and Information Technology,

College of Science, Engineering, and Health,

RMIT University,

Melbourne, Victoria, Australia.

March, 2013

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Hilal Ali Abdullah Al Maqbali

School of Computer Science and Information Technology

RMIT University

March, 2013

Acknowledgments

First of all, I would like to express my sincerest gratitude to my supervisors Falk Scholer, James A. Thom and Mingfang Wu. My primary supervisor, Falk Scholer, provided all the support needed to succeed in the world of academia. Thanks also goes to my second supervisor, James A. Thom, for his knowledge and support in developing my work. My third supervisor, MingFang Wu, aided me in my work and always provided friendly support and help.

I would like also to express my enormous gratitude to Elizabeth Beaton for her professional help in editing my thesis.

It is clear that I would not have reached this stage without the continued personal and financial support of my father and mother, they have my eternal gratitude. My wife, Miad Al-Shezawi, has been my constant source of strength and inspiration. Also, my thanks go to my friends, Qasim Al-Mamari, Shahid Al-Balushi and Lorena Leal, for their valuable advice, support and encouragement.

Finally, I would like to thank my scholarship provider, The Ministry of Higher Education in Oman, for giving me this opportunity to extend my knowledge and career. I look forward to contributing to the future of education in Oman.

Credits

Portions of the material in this thesis have previously appeared in the following publications:

- H. Ali [Al Maqbali], F. Scholer, J. A. Thom, and M. Wu. User interaction with novel web search interfaces. In *Proceedings of the 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7*, pages 301–304. ACM, 2009

- H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Do users find looking at text more useful than visual representations? a comparison of three search result interfaces. *In Proceedings of the Australasian Document Computing Symposium*, pages 35–42, 2009

- H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Evaluating the effectiveness of visual summaries for web search. *The Fifteen Australasian Document Computing Symposium (ADCS2010), Melbourne, Australia*, pages 36–43, 2010

- H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Tracking the impact of visual summaries on navigational web search. *Manuscript submitted for publication to JASIST (Under revision to address reviewers' comments)*, 2012

Note

Unless otherwise stated, all fractional results have been rounded to the displayed number of decimal figures.

Contents

List of Figures

List of Tables

Abstract

Search result organization and presentation is an important component of a web search system, and can have a substantial impact on the ability of users to find useful information. Most web search result interfaces include textual information, including for example the document title, URL, and a short query-biased summary of the content. Recent studies have developed various novel visual summaries, aiming to improve the effectiveness of search results. In this thesis, the impact and efficacy of presenting additional visual summaries are investigated through a series of four studies. User interaction with the search results was captured using eye tracking data.

In the first study we compare the effectiveness of three publicly available search interfaces for supporting navigational search tasks. The three interfaces varied primarily in the proportion of visual versus textual cues that were used to display a search result. Our analysis shows that users' search completion time varies greatly among interfaces, and an appropriate combination of textual and visual information leads to the shortest search completion time and the least number of wrong answers. Another outcome of this experiment is the identification of factors that should be accounted for in subsequent, more controlled, experiments

with visual summaries, including the size of the visual summaries and interface design. An understanding of the features and limitations of the eye tracker, particularly for IR studies, was also obtained.

To obtain a richer understanding of a user's information seeking strategies and the impact of presenting additional visual summaries, five interfaces were designed: text-only, thumbnail, image, tag and visual snippet. In the second study, fifty participants carried out searches on five informational topics, using the five different interfaces. Findings show that visual summaries significantly impact on the behaviour of users, but not on their performance when predicting the relevance of answer resources. In the third study, fifty participants carried out five navigational topics using the five different interfaces. The results show that apart from the salient image interface, users perform statistically significantly better in terms of time required and effort required to answer given navigational search topics when additional visual summaries are presented. The fourth study was conducted with both navigational and informational topics, for a more detailed comparison between the best-performing interfaces identified in the previous studies: salient images for informational searches, and thumbnails for navigational searches. The findings confirm our previous results. Overall, the salient image interface can significantly increase user performance with informational topics, while thumbnails can help users to predict relevant answers, in a significantly shorter time, with navigational search topics.

Chapter 1

Introduction

A web search engine is an essential tool that enables users to find desired information on the World Wide Web. Presentation of search results is a fundamental part of a web search engine, as the visual layout influences the way that users retrieve data and guides them towards relevant information. This thesis investigates the effectiveness and impact of additional visual summaries in search results, and evaluates the effect of such visual summaries on user searching behaviour and performance, using different topic types.

Traditional search results are typically presented as a vertical list of document summaries, where each item consists of a web page title, a short text extract from the source document, and the URL. Effective presentation can play an important role in facilitating information seeking. Previous studies show that providing additional features such as images and short text summaries in search results can have a positive influence on user performance and learning. For example, Mayer et al. [1996] found that users require less cognitive processing to understand a topic when a related visual summary is presented along with the text. The

time that a person spends digesting the meaning of a picture is enough to read only one to four words [Coltheart, 1999]. Furthermore, pictures have a positive critical impact on people's perception [Goldberg, 1991]: even users who are highly verbal and prefer text find pictures easy to understand [Mendelson, 2004]. Presenting visual summaries alongside text summaries is thus a promising method for improving users' relevance prediction ability.

Studying the impact of additional visual summaries on user seeking behaviour is essential for investigating their effective use in web search interfaces. Users employ a variety of strategies when they browse a search results page, and understanding information seeking behaviour patterns will enable interfaces to be better designed to improve user performance and cognitive processes. The impact of visually designed search results pages – particularly on user performance, browsing strategies, and interaction with informative components of the results screen – has rarely been investigated in detail.

In this chapter we define the key concepts investigated in the research questions. Section 1.1 discusses the importance of studying the effectiveness of visual summaries and Section 1.2 describes online query types. We introduce eye tracking in Section 1.3, and in Section 1.4 we describe the research questions of this thesis. The contributions and organisation of the thesis are detailed in Section 1.5.

1.1 Visual summaries

In this thesis, we use the term visual summary to refer to a graphical representation of a retrieved web page. The visual summary can be a miniature version of the retrieved web page such as a thumbnail, or a relevant image to the user's query, such as a salient image

from the underlying document. Usually, it aims to capture key identifying features of the document.

Consider the query "`Sun`". This query could be a navigational query (that is, a query intended to find a specific website) with an intended target such as the British newspaper "`The Sun`", or the Australian newspaper "`Herald Sun`". Additionally, it has several potential purposes to function as a query to find more information about specific manufactures such as: "`Sun Microsystems`", a former computer company, "`Sun Studio`", a popular recording studio, or "`Sun Ringle`", a manufacturer of wheels and other bicycle components. Furthermore, this query could be informational (that is, a query that represents an information need to learn about a topic), representing the need to find out more about the sun or about solar systems. As these examples relate to different information domains, popular search engines can provide better quality results for the query once a user enters more text in the search box. However, the additional text requires the user to possess better understanding of the query. One solution that can help users to find potential relevant answers more quickly is the addition of visual summaries.

Visual summaries have been used heavily in advertisements as they are not only attractive but can also be viewed at a glance, and are easily interpreted [Messaris, 1996; McQuarrie and Phillips, 2005]. Popular search engines such as Google and Bing also use such summaries heavily in their news search results lists. Furthermore, visual summaries can help users to reformulate queries and effectively re-find webpages [Joho and Jose, 2006].

In this thesis, four visual summaries are evaluated: thumbnails, visual tags (a novel combination of a thumbnail and a tag cloud), excerpt images (a dominant image from an

underlying document that is relevant to a user's query) and visual snippets (a combination of a logo and a salient image). We evaluate the effectiveness of these visual summaries for web search results and their impact on user seeking behaviour.

1.2 Web search queries

A query consists of one or more terms, each a string of alphanumeric characters. Web queries are commonly classified as *informational*, *transactional* or *navigational* [Broder, 2002; Rose and Levinson, 2004; Song et al., 2009; Jansen and Booth, 2010]. Based on log analysis or survey results, studies show variation in the reported percentages of query types [Broder, 2002; Rose and Levinson, 2004; Jansen et al., 2008], possibly due to the size of samples [Jansen et al., 2008], or the different definitions [Lewandowski, 2006]. However, all studies agree that informational queries get the highest percentage in comparison with navigational and transactional queries.

Informational queries are typically broad topics, where relevant answers might be found in thousands of web sources and users might need more than one page to satisfy their information need. In informational queries, the user does not target a particular website to retrieve the desired information. In contrast, *navigational* queries occur where the user is searching to find a specific single website for a given topic. For example, finding the homepage of the Microsoft Network website (MSN) is not the same as finding the homepage of the MSN games website: the latter is the hub page for a prime sub-part of the overall MSN website.

1.3 Evaluation of web search interfaces using an eye tracker

User evaluation allows researchers to get a richer understanding and characterize insights into the differences between good and poor web search interfaces. Additionally, user evaluation is essential to verify and validate novel techniques for search result representation. There are numerous well-defined methods that can be used to evaluate web search interfaces, such as think-aloud, taking notes, logging and stimulated recall. Eye tracking is a modern technique employed in usability studies that can provide useful information, such as the time participants spend viewing a particular element on the interface, participant scan paths, and task completion time.

For the user studies in this thesis, we collected experimental data unobtrusively using the Tobii T60 eye tracker. This non-invasive device captures the position of the user's gaze on the screen using infrared cameras.

1.4 Research aims and contributions

The general objectives of this study are: (1) to analyse and draw conclusions about the possible benefits of using eye tracking in information retrieval studies; (2) to evaluate the impact of additional visual summaries in search results on user searching behaviour and performance; and (3) to evaluate the impact of topic types on the effectiveness of visual summaries. In this thesis, we treat text summaries as an essential part of the web search interface, so in our analysis, user interaction with text summary components (title, short snippet and URL) is also investigated.

Figure 1.1 illustrates the overall structure of this thesis, and how we addressed these

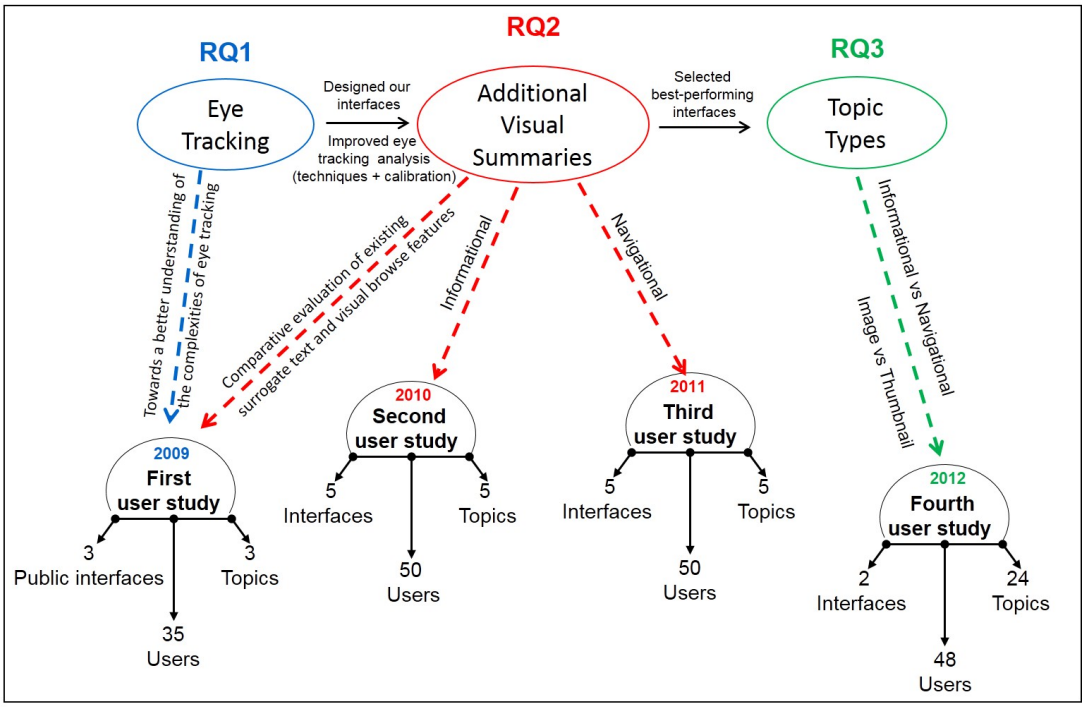

Figure 1.1: Overview of thesis structure, research questions, and key experiments.

three aims. Four user studies were conducted, where each study investigated derived sub-questions related to these aims. The first study, of *publicly available search engines*, analysed three web search interfaces that make use of different features including text summaries, clustering of information, and visual thumbnail images. For the second *informational* study, five interfaces were designed: a text-only interface and four visual interfaces. Users were asked to carry out a series of five informational search topics. In contrast, in the third *navigational* user study, the same five interfaces were evaluated using navigational search topics. In the final *comparison* study, we evaluated the effectiveness of the two best-performing interfaces from the our previous studies over a larger set of both informational and navigational search topics.

The following research questions are addressed in this thesis:

- **RQ1: Eye tracking.** How can an eye tracker be used to understand user behaviour when interacting with textual and visual summaries of search results?

  This question is addressed in Chapter 3. It is clear that when applying eye tracking for the evaluation of web search interfaces, the analysis of eye movements and the interpretation of users' interaction with search results presents researchers with many complexities. In recognition of this difficulty, we started our work with a study of *publicly available search engines* to better understand the complexities. By understanding aspects of eye tracking such as calibration and the impact of design features, researchers can better ensure the accuracy of their data. This increase in quality of data recording is consequently used to develop techniques to improve the evaluation of web search interfaces and to optimise the quality of collected data. In addition, we identify eye movement metrics for online searching (and proposed metrics) and the limitations of eye tracking.

- **RQ2: Visual summaries.** Does providing additional visual summaries for the presentation of web search results impact on users' information-searching behaviour and performance?

  This question was investigated in two stages. First, in the *publicly available search engines* study (described in Chapter 4), we evaluated three existing search result interfaces that differ primarily in their use of textual summaries and visual browsing features. In particular, we addressed two aspects of the second research question: how different search interface features impact on users' information-searching behaviour,

and how visual representations compare with text summaries in terms of effectiveness. Our analysis indicates that most users spend a substantially larger proportion of time looking at text information, and that those interfaces that focus on text-based representations of document content tend to lead to faster task completion times for named-page finding searches. However, an appropriate combination of textual and visual information leads to the shortest search completion time and the least number of wrong answers.

In the second stage, based on the findings of the *first study*, we designed five web search interfaces to evaluate different approaches to visual summaries. Each interface presents exactly the same text summary, but with different visual summaries (except for a text-only interface consisting entirely of text summaries). The following sub-questions were then addressed in the *informational* (Chapter 5) and *navigational* (Chapter 6) studies: how do additional visual summaries influence the ability of users to predict relevant answers, task time completion, user interaction with the results screen, and mental effort expended ?

In the *informational* user study, the results show that visual summaries significantly impact on the behavior of users, but not on their performance when predicting the relevance of answer resources. Users spend significantly less time looking at the textual components of summaries when additional visual summaries are presented. Comparing the performance of users in predicting the relevance of answer pages with a text interface versus visual interfaces suggests that the tested visual summaries can mislead users to select non-relevant items on informational search topics, although this difference was

not significant.

In the *navigational* user study, an additional investigation was conducted into user browsing strategies to learn more about user interactions with the presented informative components. The results show that for particular types of visual summaries, users perform significantly better in terms of time and effort required to complete search tasks when additional visual summaries are presented. Comparing the various amounts of attention spent on visual summaries suggests that the more time that a user spends on these summaries, the more their ability to correctly predict the relevance of answers is increased. Less effort and time are required to find the answer. Moreover, different amounts of attention spent looking at the additional visual summaries indicate different forms of browsing.

- **RQ3: Topic types.** How does the type of search topic influence the effectiveness of additional visual summaries for the presentation of web search results?

  Based on the findings of the *informational* (in Chapter 5) and *navigational* (in Chapter 6) user studies, the best-performing visual interfaces for each type were selected (image for informational and thumbnail for navigational). In the *comparison study* (described in Chapter 7), twenty-four topics (12 informational and 12 navigational) were used to evaluate the impact of the topic types on the effectiveness of the two chosen interfaces. We evaluate the impact of topic types and visual summaries (thumbnails and images) on user search behaviour and performance. We also study user interaction with specific text summary components (title, short snippet and URL) in detail.

The results confirm our previous findings: users perform significantly better with informational topics when using the salient image interface. In contrast, for navigational searches, the thumbnail interface not only helps users to find relevant answers in a shorter amount of time, but also requires significantly less user effort.

1.5 Thesis organisation

This thesis presents four user studies to investigate the research questions discussed in the previous section. Each user study is discussed in a separate chapter, where the experimental setup of the study and involved interfaces are described. The thesis is organised as follows:

- **Chapter 2** begins by presenting a general introduction to information retrieval and the traditional evaluation measures of web search interfaces. A review of the literature on the effectiveness of visual summaries is then presented, and different approaches to visual summaries are discussed.

- **Chapter 3** begins with a description of eye tracking, and explains different types of eye tracking. Definitions of common eye movement measures and our proposed metrics are then provided. Limitations and difficulties of eye tracking metrics are also discussed. We introduce the methods that were followed when analysing data collected in our user studies.

- **Chapter 4** describes the first experiment, which analyses three publicly available search interfaces and compares the effectiveness of textual information and additional browsing features. The three interfaces vary primarily in the proportion of visual versus

textual cues that are used to display a search result.

- **Chapter 5** describes the second experiment, which evaluates the effectiveness of presenting additional visual summaries for informational topics.

- **Chapter 6** describes the third experiment, which evaluates the impact of the presence of additional visual summaries on user's search performance and behaviour with navigational topics.

- **Chapter 7** describes our fourth study, conducted to compare between the two best-performing visual interfaces according to our previous studies (salient image for informational topics, and thumbnails for navigational topics) across mixed tasks and a larger number of topics.

- **Chapter 8** presents the conclusions of our research and discusses potential future work arising from our findings.

Chapter 2

Evaluation of visual summaries in web search interfaces

A search engine is one of the most powerful tools used to access the World Wide Web, and the presentation of search results is a fundamental component of a search engine. The presentation and organisation of results should allow users to gain a better understanding of the data and obtain easy access to the retrieved documents. Unfortunately, providing well-organised search results is not always simple, as users' needs are complicated and not easy to predict.

The focus of this thesis is the evaluation of the effectiveness of different methods of presenting search results. Thus, we begin in Section 2.1 with an overview of the different methods and techniques for evaluating web search interfaces. In Section 2.2 we describe the current presentation of search results, identifying problematic issues related to text summaries and suggesting alternative approaches employing visual information. This thesis also seeks

to investigate the effectiveness and impact of additional visual summaries on user search behaviour. Therefore, in Section 2.3 we describe the importance of visual summaries, while Section 2.4 introduces several approaches to visual summaries. Section 2.5 describes some of the effects of the use of visual summaries in web searching. Section 2.6 outlines the current challenges of presenting visual summaries for web search results, and a summary of the chapter is presented in Section 2.7.

2.1 Evaluation of web search interfaces

Interactions between users and information retrieval systems can be grouped into three categories: query formulation, relevance prediction and relevance evaluation [Dziadosz and Chandrasekar, 2002]. Both the relevance prediction and relevance evaluation categories involve relevance assessment of web pages. In relevance prediction the assessment is based on a representation of the web page, while in relevance evaluation, the assessment is based on the actual web page [Carol, 1998]. Various methods have been introduced to evaluate user performance and usability, where the following sections describe some of these methods.

2.1.1 Conventional usability methods

Usability can be determined by two broad types of approaches: usability inspection, and usability testing. Usability inspection is a set of methods used to evaluate a user interface, relying on expert inspection [Nielsen, 1994]. In contrast, usability testing focuses on measuring the efficacy of the interface through the performance of real users.

Many traditional methods have been used to collect data in IR user studies, such as the

observation of user behaviour, self-reporting, questionnaires, interview, and think-aloud. The use of questionnaires is the most popular quantitative method employed to obtain various aspects of user behaviour, such as subjective reactions to the use of interfaces. Questionnaires can be determined by scale (for example, a three or five point scale) or can be open, where users can type their responses with no restriction. However, one problem arising from the use of questionnaires is that they are not as flexible as other methods, since the questions remain fixed.

The observation of user behaviour, (for example, observing the time required to complete a task and make navigational choices) can be conducted by one or more protocols such as paper and pencil, video or audio recording, or both. These observation protocols are easy to implement but difficult to use to extract information; they require good skills to transcribe user interaction and avoid incorrect results [Shneiderman and Plaisant, 2006]. In self-reporting, users describe the steps they took whilst browsing a given system. In think-aloud methods, users are asked to verbalise what they are thinking about and describe their decision making [Ericsson and Simon, 1985]. Self-reporting and think-aloud protocols are subjective and require little expertise to administer [Liou, 2000; Masarakal, 2010].

Interviews are also a popular protocol in usability studies, allowing a variety of question levels, these usually start with general discussion and end by discussing specific issues. However, the personality and character of interviewers and participants can impact on the responses [Fidel, 1993]. In addition, responses to the interview questions can be interpreted subjectively [Ericsson and Simon, 1985].

2.1.2 Techniques employed in formal studies

Hoeber [2006] outlined five methods to evaluate the use of visual web search interfaces, namely inspection, laboratory, field trails, longitudinal, and instrumentation and log analysis. A lab study is the classic name given to formal usability studies, because these studies are conducted in a usability laboratory or an appropriate room with chair, desk and computer [Hearst, 2009]. Different techniques can be used to observe user behaviour such as taking notes, recording (video and audio), or capturing the screen.

Analysis of log files is a technique used to capture user browsing behaviour. Logs can be instructed to record a wealth of information about user behaviour when users are interacting with a system. For example, a typical log might record the browsing of a specific document set. Click-through data is used to evaluate user performance by collecting the data provided in the log file. Researchers can analyse specific events, times and internal navigation to understand different aspects of user seeking behaviour [Bowman et al., 2001]. Click-through data has been used in various studies, such as search ranking [Zhu and Mishne, 2009], query clustering [Li et al., 2008a], search personalization [Teevan et al., 2008] and search difficulty prediction [Mei and Church, 2008].

A tool can also be used to provide statistical summaries of a user's search behaviour, such as SpeedTracer, developed by Wu et al. [1998]. SpeedTracer provides summaries with reports about user navigational paths, the most common events and frequently visited pages. However, click-through data cannot provide information on how much attention was spent on particular targets on the web page. In addition, click-through data is effective for a large set of data where the user is instructed to browse web pages to answer the given tasks. We

used a log file in one of our user studies to track users' task completion time.

Another modern technique employed in usability studies is the tracking of eye movements. Several studies show that eye tracking is an accurate means of evaluating different interface designs [Radach et al., 2003; Lin and Zhang, 2003; Xu et al., 2008]. Various studies used eye tracking to evaluate different components of search interfaces [Granka et al., 2004; Rele and Duchowski, 2005; Cutrell and Guan, 2007; Eger et al., 2007; Guan and Cutrell, 2007], however, a few papers discussed how an eye tracker can be used in the evaluation of alternative web search interfaces. In this thesis, we used eye tracker to capture user' eye movements to evaluate the effectiveness of presented information. More details of eye tracking methodology and descriptions are given in Chapter 3. In addition, we used questionnaires to collect feedback on particular aspects of our user studies.

2.1.3 Traditional information retrieval measures

Performance evaluation should include both users and systems, particularly for web search interfaces, and evaluation should consider system functionalities and error analysis [Baeza-Yates and Ribeiro-Neto, 1999]. Evaluation of user performance requires consideration of the search task (task scenario) and presented information (presentation and organisation). Standard metrics in IR take into account different variables (depending on the selected metric), such as response time, total number of relevant retrieved results, and selected items. In this thesis, the effectiveness of visual summaries is also measured by Click Precision, Click Recall, and Click F-measure.

Click Precision measures the correctly identified relevant answers as a proportion of

all answers that the user selected. In other words, Click Precision measures how successfully users were able to find relevant answers for a given search task, where credit is given for accurately selecting relevant items rather than for the total number of items.

$$\text{Click Precision} = \frac{Total\ relevant\ selected\ answers}{Total\ selected\ answers}$$

Click Recall shows the number of relevant answers selected by users as a proportion of the total number of relevant answers available for that topic. In other words, Click Recall measures the ability of user to select all available relevant answers among given items. In studies employing Click Recall, users are expected to continue searching to find all potential relevant answers to get higher score.

$$\text{Click Recall} = \frac{Total\ relevant\ selected\ answers}{Total\ available\ answers}$$

Click F-measure calculates the harmonic mean between Click Precision and click Recall. Improvement in Click F-measure is more sensitive to Click Precision than to Click Recall.

$$\text{Click F-measure} = \frac{Precision \times Recall}{Precision\ +\ Recall}$$

Other metrics *Precision at recall point* is defined the same way as Click Precision, but it measures specific document ranking numbers such as (P@5) or (P@10). It is used for a large number of presented items, where the user is likely to look at the first few items on the screen. As we did not utilise such a large group of items in our user studies, we did not use this metric. *Mean Average Precision* (MAP) is another common measure, providing the mean of the average precision of a given set of queries, while *Average Precision* computes the average value of precision across all values of recall. Average Precision was used to design the test set collection for one of our user studies to have a good spread on position and number

of relevant answers for each topic. More details are discussed in Chapter 7.

Studies show that the above metrics are good for effective retrieval. However, they cannot provide precise information on user interaction with the informative components, such as time spent on each item and number of items viewed. One goal of this thesis is to evaluate the effectiveness of additional visual summaries on web search results. Providing precise information on user interaction with the presented information allow us to gain a richer understanding of the summaries' impact on user searching behaviour. Therefore, in addition to the use of some of the above metrics, we used an eye tracker to capture the user's eye movements and gain a deeper understanding of naturalistic search behaviour.

2.2 Presentation of search results

The Presentation of search results is a fundamental component of a web search engine. The presentation of search results influences users' assimilation of the context and guides users to look for information that is relevant to them.

2.2.1 Textual summaries

Traditionally, search results are presented as a vertical list of textual summaries, where each summary consists of a web page title (typically short), a small text extracted from the source document, and the URL [Tombros and Sanderson, 1998; Wu et al., 2001]. This combination of three elements is expected to give a user an overview of the actual web document, based on the user's query.

Cutrell and Guan [2007] evaluated how users interact with textually-presented search

results, using an eye tracker to collect experimental data. The study involved twelve tasks (6 informational and 6 navigational). Results show that users spent significantly more time viewing the top-ranked items. User browsed the search results vertically with top items receiving more attention and being viewed earlier than the ones further down the list. The results also showed that users spent more time looking at the title and snippet than at the URL.

One recent study, from the middle of 2000, conducted a survey by interviewing 566 adults over the phone. The results show that only 21% of users found relevant results when querying a search engine, and that 75% were disappointed with the results returned, while 4% did not answer [Sullivan, 2001]. In addition, 89% of the participants felt that search engines could be improved, this defines a major challenge in investigating and solving missing information. The way in which users interact with the search result interface may be one factor contributing to a poor user experience. Furthermore, some issues with textual summaries have arisen to hinder users' online searching. Some of these issues are discussed below.

Search engine spam

A search engine's goal is to find better or trusted word content on the web for a given string query, using algorithms. Spams are websites that manipulate one or more of the pre-defined processes or rules to make them seem more relevant [Sullivan, 2008]. The goal of spam is to be shown (higher in the ranking of search results) as relevant information despite the fact that, upon visiting the actual page, it turns out to be irrelevant. Spam websites can employ various techniques to present themselves as relevant, such as repeating links of

other websites, increasing the frequency of query terms, or copying content. Gyongyi and Garcia-Molina [Gyongyi and Garcia-Molina, 2005] classified the spamming techniques into nine techniques, all of which use text to increase the website's ranking on the search results list. Presenting additional visual summaries of search results may decrease the impact of spam: although visual spam exists, it is not as frequently presented as text.

Ambiguous queries

Many web pages are retrieved by search engines through the use of a query term, and users must evaluate presented information or reformulate their query to find potentially relevant answers. If the query is ambiguous, then this process becomes more challenging and complicated [Smeaton, 1992]. Ambiguity in queries has been studied widely, where the lexical ambiguity of a query can be caused by a short query strings that might have several potential meanings [Krovetz and Croft, 1992; Voorhees, 1993; Sanderson, 1994; Sanderson and Van Rijsbergen, 1999].

The shorter the query is, the more likely that it is an ambiguous query. Jansen et al. [1998] and Sanderson [1999] report that about 17-23% of logged queries are ambiguous based on query length (average length is close to one sting). For example the query term "`Sun`" is highly ambiguous and may serve several potential purposes: to find information about the British newspaper " `The Sun`", the Australian newspaper " `Herald Sun`", " `Sun Microsystems`" or other terms as discussed in Section 1.1.

2.2.2 Visual information

The visual presentation of web search results can take different forms, with a wide range of features. The main goal of visual information is to enable individuals to use and understand the retrieved documents more easily. Previous studies show that the provision of additional features in the presentation of search results, such as displaying visual features along with the short text summaries, can have a positive influence on user performance [Ayers and Stasko, 1995; Ogden et al., 1998; Sutcliffe et al., 2000; Kaasten et al., 2002; Dziadosz and Chandrasekar, 2002] and on learning [Waddill and McDaniel, 1992]. Many techniques have been proposed: some have already been implemented in existing search engines, while others have only been recommended by researchers. Highlighting and colouring query terms, clustering, background colouring, zooming, altering the size font of the query terms and using different font colours are good examples of visual techniques that can be used to organise and present web search results. In this section, some examples of visual search techniques are discussed.

Clustering

Chen and Dumais [2000] used ambiguous queries to evaluate user performance using well-defined clusters. Results show that users were able to locate relevant answers much more quickly than in a traditional search results list. However, one major issue with clustering is that the clusters number or label must be known in advance [Krishnapuram et al., 1995].

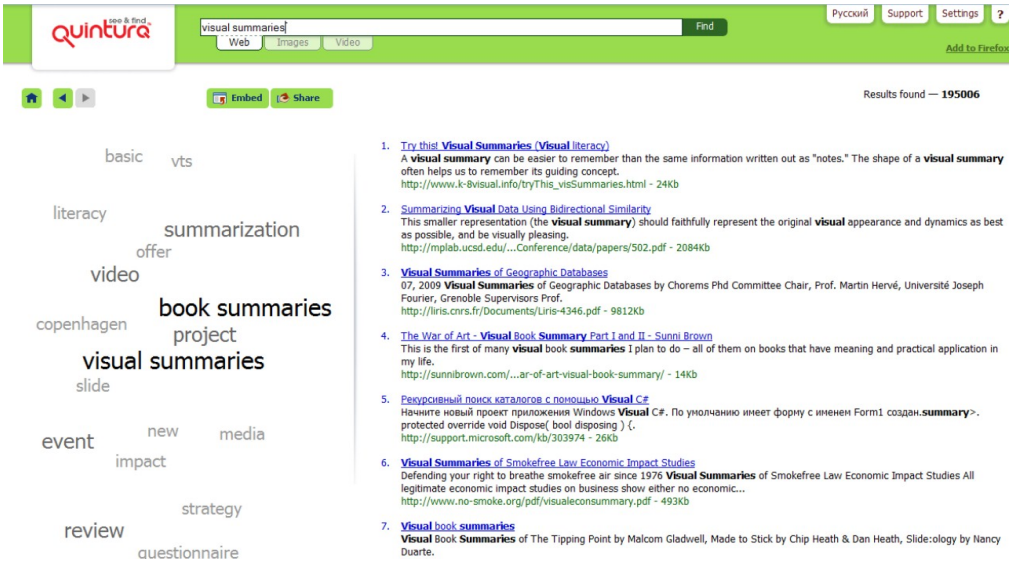

Figure 2.1: Search results for the query "visual summaries" using Quintura *search engine.*

Cloud tag

The Quintura search engine uses the cloud technique that shows related (suggested) words for the query, along with a search results list (see Figure 2.1). If a user clicks on a word in the cloud window, then that word will be added to the existing query terms and search results will be refreshed.

TileBars

Hearst [1995] developed the TibleBars interface for web search results, where the interface presents a visual summary (called TileBars) along with a short text document surrogate for each item. TileBars is a visual approach that processes the document content and length in order to present a set of bars based on frequency of query terms and document length. It allows the user to gain an indication of the retrieved document length and the query frequency without visiting the actual document. Hoeber and Yang [2006] also developed a

web search interface using TileBars.

Visual summaries

In this thesis, the term visual summary refers to a graphical representation of a retrieved web page. Marsh and White [2003] investigate the relationships between images and text for 945 images-text pairs that were retrieved from educational, newspaper and business web pages. The results show that the functions of images can be classified into three groups based on their relationship to the text: little relationship to the text, close relationship, and based on (but overriding) the text. This classification emphasises the potential usefulness of including additional visual summaries on web search results. One indication is that visual data on the World Wide Web (such as images and a web site's visual layout) has a strong relation to the provided text and hence can lead to a positive impact on user understanding and comprehension of the context. Thus, presenting a whole or a composite part of a related visual object along with the document surrogate can have a similarly positive impact on user performance.

One of the popular visual summaries is the thumbnail (a screen shot of the retrieved web page) see Figure 2.2. However, news search engines such as the news search features on Google and Yahoo! usually present the dominant image of the article. In addition, researchers have proposed other approaches for visual summaries, which will be discussed in Section 2.4.

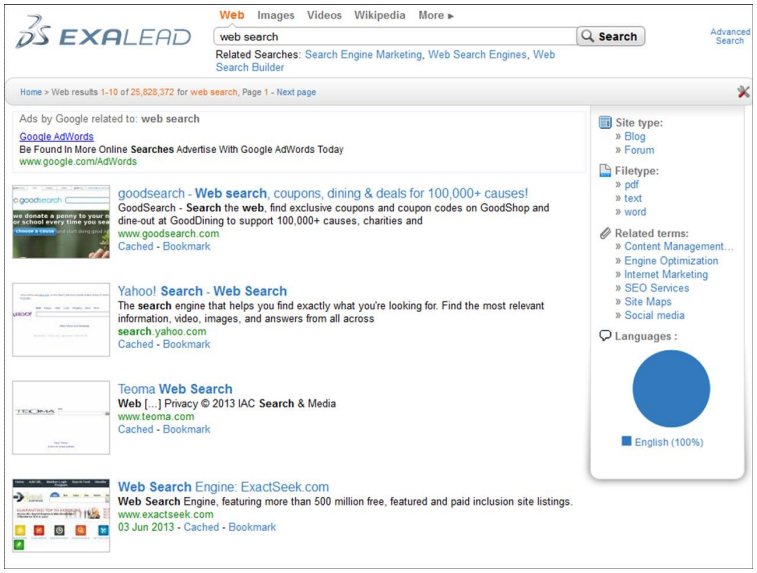

Figure 2.2: Search results showing thumbnails for the query "web search" using the Exalead *search engine.*

2.3 Visual summaries are promising

Visual summaries (thumbnails) have been used in different fields for a variety of purposes. For instance, thumbnails were recommended and used in many developed browsers using different types of presentation such as a (3D) data mountain [Maarten et al., 1999], hierarchy nodes [Hightower et al., 1998], zooming [Jhaveri and Räihä, 2005] and others such as a vertical list [Ayers and Stasko, 1995; Cockburn et al., 1999; Won et al., 2009].

Lam and Baudisch [2005] proposed a technique called *Summary Thumbnails* for small screen web browsers such as mobile phones. Summary Thumbnails generates an enhanced thumbnail with readable text parts (mostly headlines) and distinguished visual features of the web page. The readable text fragments in *Summary Thumbnails* provide the user with a hint of the article under that link, and Lam and Baudisch [2005] suggest that this technique helps users to browse with no need to zoom. Qualitative and quantitative user studies were

conducted where Lam and Baudisch evaluated the effectiveness of their *Summary Thumbnail* interface by comparing it with traditional thumbnail and single-column interfaces (showing text only, with no images). In the qualitative user study, nine participants were asked to browse the BBC news website using one of the previous three interfaces to read an article (based on their interests). Participants were instructed to think aloud during browsing, then an interview and a questionnaire were conducted. In the quantitative study, eleven participants were asked to answer nine tasks followed by a short questionnaire. Results show that *Summary Thumbnails* produces a better performance by enabling users to reach their goals much faster than other interfaces and with lower error rates.

Furthermore, Outing [2004] found that online users are impatient and prefer reading short articles, so additional visual summaries can support this behaviour. Presenting visual summaries might help to deal with this issue, as a picture is often seen to be worth a thousand words. In addition, studies show that images make reading more interesting and attractive [Peeck, 1993; Mayer et al., 1996; Pekta, 2012].

Cockburn et al. [2006] developed a new tool called Space-Filling Thumbnails (SFT) to improve the navigation of digital documents using a scrolling technique. The interface of SFT provides two views for documents: (1) all pages shown as thumbnails in one non-scrolling page (a small size of 34×44 pixels is used for a large number of pages), and (2) a dynamic view for selected pages with a size of 154×200 pixels (the user is required to click on pages to enlarge the view). Three experiments were done to evaluate the SFT by comparing it with other candidate systems, examining different aspects such as visual scan and spatial memory. For example, to evaluate spatial memory, each participant was given a

document containing thirty pages and asked to conduct two tasks: (1) finding specific pages in the given document, and then (2) repeating the same task to test the impact of images on spatial memory. Cockburn et al. [2006] used questionnaire responses (5-point Likert scale), document length and task completion time to measure user performance and the effectiveness of thumbnails. One of the results indicates that small thumbnails provide a good cue for page layout. Findings also suggest that images improve the user's ability to visually scan and have a positive impact on spatial memory (making it easier to recognise the target location). SFT provides good results when compared with other candidate systems.

2.4 Approaches to visual summaries

In this study we focus on visual summaries that present information graphically. Several types of visual summaries have been developed. A thumbnail (miniature image of a web page) is the most popular visual summary and has been examined in many studies. Thumbnails are already used in some existing search engines, as will be shown later in this thesis.

In the past, quite a few studies explored the visual presentation of search results [Shneiderman, 1996; Card et al., 1999; Shneiderman, 2008]. A proper visual representation can communicate some kinds of information much more rapidly and effectively than textual representation. Several novel approaches have been developed for visual summaries to improve user performance in finding desired information: some of these approaches use the snapshot of a web page, such as an enhanced thumbnail [Woodruff et al., 2001], whereas others use a salient picture within the retrieved web page such as a salient image [Li et al., 2008b] or visual snippet [Teevan et al., 2009]. More details are discussed below.

2.4.1 Enhanced thumbnail

An enhanced thumbnail, proposed by Woodruff et al. [2001], is another visual summary that consists of a thumbnail with highlighting and enlarging of the queried terms. Woodruff et al. [2001] evaluated these enhanced thumbnails by comparing them with text-only summaries and plain thumbnails. The search results list presented only one of the three summaries for the given search topics. Four task categories were involved (locating a picture, a homepage, shopping, and an informational query). The results showed that the category of the task has a strong effect on user performance. However, an enhanced thumbnail summary showed only statistically indistinguishable effects on user performance compared with text-only and plain thumbnail summaries.

2.4.2 Salient image

Li et al. [2008b] crawled three websites (MSN.com, MIT.edu and CNN.com) to evaluate an interface using image excerpts (relevant dominant images for the user's query along with text summaries) compared with a text-only interface. An algorithm was used to collect the image excerpts and order them according to their scores on relevance to the query so as to present the highest scoring image excerpt for the given query. A meta search engine was built for the image excerpt interface, based on a search results list from Google. Google results were used to obtain the text-only summaries in their study. The study involved two types of query: informational and navigational. The results showed that the image excerpt together with a text summary helped the user to find answers in less time than the text-only interface. Loumakis and Grayson [2011] compared three interfaces: text-only, image excerpt only and

a combination of images and text cues. The results showed that the images can provide cues for users, but not as well as text summaries. Moreover, users showed a preference for images: indeed, the researchers found that although image information can be interpreted in different ways by users, presenting images along with text can outweigh the use of text alone.

2.4.3 Visual snippet

Teevan et al. [2009] developed visual snippets that consist of three components: a page title, a salient image and a logo of a website. The salient image is the most relevant image from the source document for the given query. If the salient image is not available, a snapshot of the web page is taken instead, while the logo is not included if it cannot be identified. Teevan et al. [2009] evaluate three types of web page representation: visual snippets, text-only and plain thumbnail. Each participant undertook twelve tasks where each group of four tasks represented a different category (homepage, shopping and medical information) using different types of summaries for each of the four tasks. The results showed that visual summaries allow users to view the retrieved web page without visiting the actual web page. Furthermore, results suggest that visual summaries show some improvement in user performance for re-finding pages that have been previously seen. However, the results show no significant improvement in the time required to answer the given informational queries when using visual snippets and thumbnails compared with text-only summaries. The participants' subjective preferences were significantly higher for visual snippets and text-only summaries when compared with thumbnails.

2.4.4 Visual summaries in existing search engines

While many search engines primarily show text-based summaries (such as a page title, a short textual snippet, and a URL), some existing search engines also provide visual features. One of the most common types of visual feature is the visual summary, such as a thumbnail (mostly used in web search results) or dominant image (mostly used in news search results). Hearst [2009] strongly recommends avoiding complexity in the presentation of search results. This recommendation explains why popular search engines such as Google and Yahoo keep their web search interfaces simple: although they have historically focused more on improving textual summaries for each result page, they have started to show simple visual summaries for some of the top search results. Other search engines, such as Oolone[1], Nexplore[2] and Search-cube[3], display a visual summary for every result in their answer list.

Different techniques are used to present thumbnails; some interfaces provide a separate column to view thumbnails for each retrieved web document[4], whereas other interfaces display thumbnails on more than half of their screen[5]. In addition, some interfaces show the thumbnail only when the mouse moves over a text abstract related to a retrieved document such Google, Bing and others[6], while other interfaces[7] show large thumbnails (using a slide technique to browse the results) with a shorter text summary in comparison with tradition textual search results. Although various different techniques of presenting thumbnails in web search interfaces exist, few of these have been tested [Woodruff et al., 2001; Xue et al., 2006;

[1] www.oolone.com
[2] www.nexplore.com
[3] www.search-cube.com
[4] www.nexplore.com
[5] www.middlespot.com
[6] www.ziipa.com
[7] www.redzee.com

Joho and Jose, 2008; Xu et al., 2009], and where testing has been performed, analysis focusing on the impact of visual summaries on user search behaviour was limited. Most studies focus their analysis on task completion time and error rate, however a few of them investigate user searching behaviour such as user effort and user interaction with the informative components. In this thesis, we look at different aspects of user searching behaviour, using an eye tracker to capture the user's eye movements on screen.

Recently, Google introduced a new tool called "Knowledge Graph" that is located on the top or the right of the screen of search results, showing narrow pictures and textual summaries for the given query [Singhal, 2012]. Knowledge Graph summarises relevant information about the topic from different sources such Wikipedia and Freebase, where additional facts are given for that particular topic. This tool mostly focuses on popular objects such as famous buildings, people, landmarks and cities. Figure 2.3 shows an example for the query "Melbourne" where the *Knowledge Graph* is located to the right of the search results list (*Knowledge Graph* can be also located on the top of the search results list for other queries).

2.5 Effectiveness of visual summaries in web searching

Studying the impact of additional visual summaries on user seeking behaviour is essential for understanding the effective use of visual summaries in a web search interface. Analysing user browsing behaviour while reviewing search result pages can provide relevant insights into how to improve user performance and cognitive processes.

In this section, we discuss most of the research related to the use of visual summaries for web search results. We classified the studies into groups based on the overall trend of the

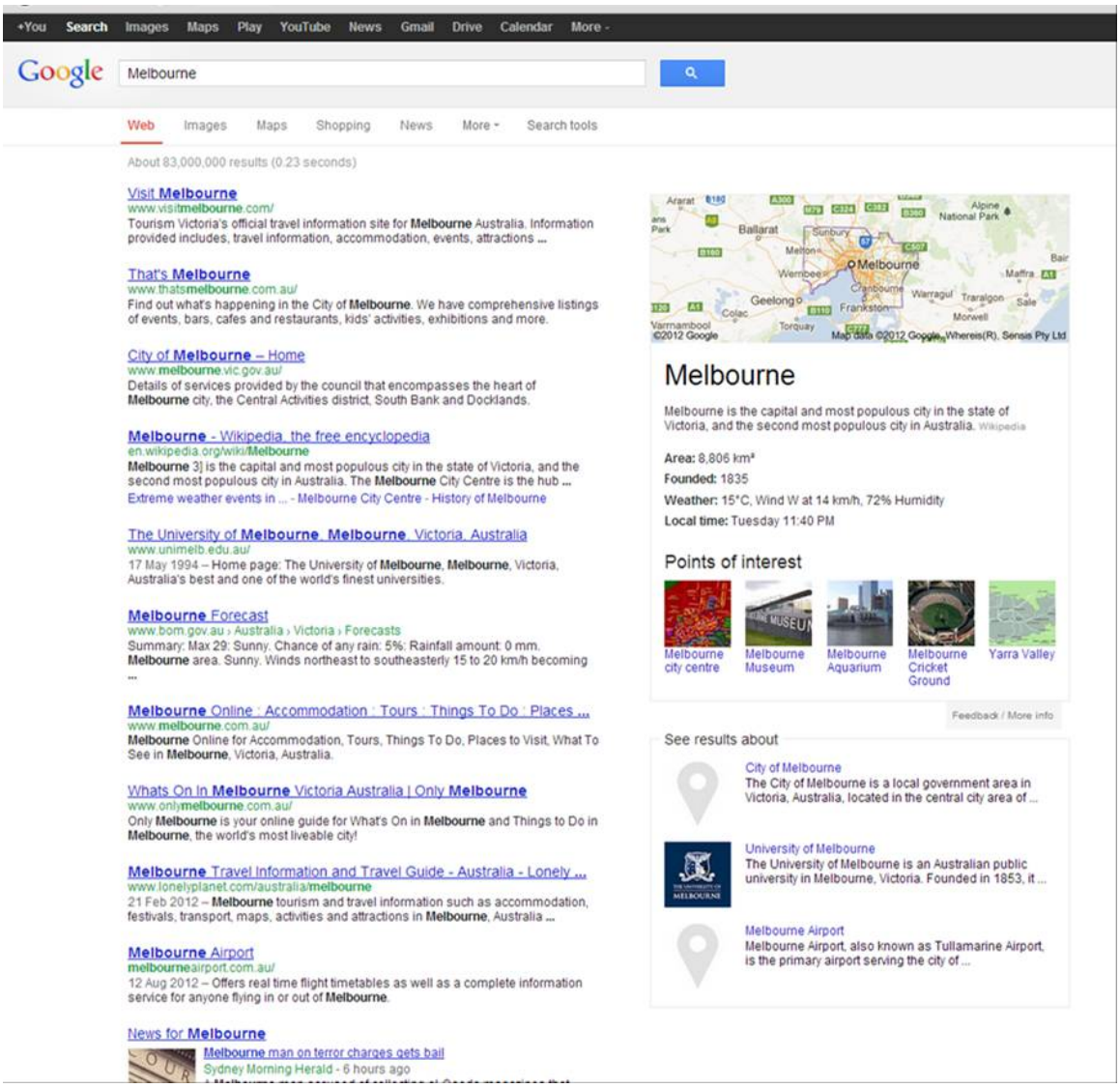

Figure 2.3: Search results for the query "Melbourne". Knowledge Graph *is provided on the right of the search results list.*

study goals on analysing user searching behaviour and effectiveness of visual summaries.

2.5.1 Finding and re-finding

Re-finding is the process where a user attempts to visit a web page that has already been visited previously. A study analysed web query logs for 114 users over one year and found that 40% of the queries involved re-finding information [Teevan et al., 2007]. Jones et al. [2002] observed user behaviour in relation to the re-use of web information. Eighteen participants with special experience were recruited from three distinguished populations: (6) managers (8) information specialists and (4) researchers. In the study, participants answered a questionnaire to provide information about their background, education and web use experience. After a few days, the following procedure was completed in an hour: an interview was conducted with the participant and then followed up with a web task customised to the participant's interest. Participants were instructed to think aloud during the web task session and a video recording was set to capture their hand movements and motion of the mouse on-screen. Ten techniques used by participants to reuse web information were observed, such as emailing information to self, email to others, bookmarking pages and saving files in the computer. Jones et al. [2002] classified these techniques into ten categories based on their functions, and results showed that users employ various techniques to record the location of interesting web pages because they find it difficult to re-find them. Thus, it is highly recommended that search results should enable users to re-find previously seen web pages, to improve effectiveness.

Re-finding information on a traditional search results page is a hard task: it requires remembering the exact phrases of the query to retrieve web pages [Teevan et al., 2004; Aula

et al., 2005]. Capra [2005] studied the impact of finding and re-finding tasks on user behaviour, and the results showed that the provision features (search engine utilities such as localised search results) help users to re-find previously seen web pages. Research indicates that thumbnails are effective for re-finding web pages that have been visited previously [Dziadosz and Chandrasekar, 2002; Do and Ruddle, 2012]. It is more difficult to remember the exact query terms than to see a snapshot of that web page and re-find a previously seen web page. In some studies, results show that thumbnails have a positive impact on helping users to re-find a previously seen web page [Ayers and Stasko, 1995; Maarten et al., 1999].

Buscher [2009] presented a new system that aimed to help users to easily recognise previously seen web pages by providing a thumbnail for each document, represented as nodes in the tree, and presenting text summaries (URL and document title) for selected documents at the bottom of the screen. The thumbnail image used in this study showed the top left corner of the original document.

A study by Yoo et al. [2008] presents a new web browser system that relies totally on thumbnails to navigate the web. The study showed that using thumbnails for browsing is effective and convenient, and helps users to easily move between pages. Thumbnails were also found to decrease the number of undesired pages visited.

Other research such as that of Jiao et al. [2010b] evaluated three visual summaries for re-finding tasks: an image excerpt from the retrieved web page, thumbnails and visual snippets proposed by Teevan et al. [2009]. However, if the internal image excerpt was not available, an external representative image was retrieved to summarise the web page. Subjects in the first part of the study were asked to type their expectations of the web page content for the

given visual summaries, then to rate their expectation after visiting the actual corresponding web page. A few hours later, in the second part of the study, the subjects were asked to re-find the web page based only on one visual summary. For the first part of the study, they classified the web pages into six categories based on amount of text, availability of dominant images and document length. Results show that thumbnails are not good summaries of pages with a large amount of text and fewer images, however, thumbnails are better than external images for websites with logos and previous seen websites. The results also show that the type of web pages and tasks impact on the effectiveness of presenting visual summaries. For example: while thumbnails are effective for simple web pages, visual snippets are more effective for re-finding web pages.

Visual summaries may therefore have a strong positive impact on re-finding previously seen web pages, and since this has been convincingly established, we did not investigate the matter further in this thesis. Our investigations were more concentrated on studying the impact of additional visual summaries on user searching behaviour in aspects such as effort and time spent.

2.5.2 Presenting additional visual summaries along with textual summaries

Dziadosz and Chandrasekar [2002] evaluated the usefulness of thumbnail images in web search results, comparing textual summaries, thumbnails, and a combination of both textual summaries and thumbnails. The results of the study showed that presenting textual summaries along with thumbnails help users to assess the potential relevance of search results to decrease the ratio of errors. Presenting thumbnail images without textual summaries limit the ability

of subjects to assess relevant material.

Czerwinski et al. [1999], conducted a study focused on the effectiveness of thumbnail images, mouse-over text and spatial location memory in 3D environment. The experiments were done in two sessions with a period of four months between them. Four types of abstract were used in this study: title only, textual summary only, thumbnails only or all the previous three items together. The data set consisted of 100 pages. During the test session, subjects were given one of the four abstract types and asked to find a related page. The results showed that thumbnail images can be valuable in recognising previously seen pages, and hence help users to find their desired information more quickly and accurately. Thumbnails halved the number of failed attempts to re-find web pages. A thumbnail image was ranked as one of the best features to help subjects to find required information.

A study by William et al. [1998] presents a system consisting of two windows. One displayed the thumbnails of the top 20 search results on the left, highlighting query term positions. The second window presented a single document with a fish-eye view, identifying the regions where query terms occur. The study compared this system with the traditional system of presenting search results, and results showed that users significantly improved their ability to identify relevant documents when using the new system. Thumbnails enabled users to scan the retrieved documents and the frequency of query term occurrences. They also allowed users to compare results in a glance. The fish-eye view allowed users to locate the relevant passages very quickly and easily.

A system to search journal articles, called BioText, was developed by Divoli et al. [2010]. BioText is an interface that uses thumbnail images to present search results. It allows users

to display search results in five ways: (1) full text with a figure display, (2) textual summary with a vertical list figure display, (3) textual summary with a grid list figure display, (4) textual summary with a table display, and (5) a detailed article summary display. The study found that (1) full text with figure display, and (2) textual summary with a vertical list figure display were the most preferable interfaces for users.

Xu et al. [2009] developed a new visual search interface for web browsing where search results are classified into groups of topics (semantically oriented). The interface was hierarchically organised and presented key messages and pictures for each topic. A simple heuristic was used to display images on the search interfaces: if the image was larger than 200×200 pixels and was not located on the corner of web page or was floating around (such as advertisements) then it was considered an actual content image. The results of the evaluation showed that the visual interface helped users to find higher quality answers, lowered completion time, lessened the total number of clicks, and satisfied users more than the traditional interface.

Joho and Jose [2006] studied the effectiveness of additional textual and visual features (thumbnails) for search results. Four interfaces were designed for this study, using Google as a base line. The first interface presented traditional text summaries (document title, short text snippet and URL), the second interface presented the same summaries as the first interface with the addition of the top ranking sentences (TRS): utility textual sentences from the retrieved document that are relevant to the given query [Tombros and Sanderson, 1998]. The third interface presented thumbnails along with traditional text summaries, while the fourth interface presented all three summaries: TRS, thumbnail and traditional document

surrogates. The results showed that the additional elements were helpful to users for assessing the relevance of results and query re-formulation. The study also showed that users' search experience plays a significant role in the facilitation provided by textual and visual features.

2.5.3 Comparing the effectiveness of different approaches to visual summaries

A study by Aulu et al. [2010] evaluated the effectiveness of presenting textual and thumbnail summaries, also aiming to understand how they help users in predicting the relevance of a web page for a specific query. The first part of the study examined two types of thumbnail: Zoomout, which is presented on 200×250 pixels, showing a larger area of the web page than a traditional thumbnail; and Zoomin, which is presented on 280×250 pixels, zooming in the top left corner of the web page. Four topic categories were used (medical, travel, shopping and images), two topics were specified for each category, and four results were generated via Google search results for each topic. The results showed that both thumbnail previews are effective, however statistical testing showed that Zoomout thumbnails provide better information. In the second part of the study, twelve participants were asked to rate (before and after visiting each web page) the helpfulness of the presentation of traditional textual summaries and the summaries containing the URL and title above each Zoomout thumbnail. The results showed that user performance was better when using thumbnails; however, this was not statistically significant. The results also showed that the location of URLs and titles affect how users browse thumbnail images, as placing text snippets below thumbnail images increases user attention. In other words, thumbnails with web search results can provide good cues and search results organisation influences user searching behaviour.

2.6 Challenges of presenting visual summaries for web search results

A rich collection of visual information is available on the World Wide Web, but popular search engines present very little of this information, considering its abundant availability. However, the visualisation of textually represented information is difficult and challenging [Hearst, 2009]. In the following sections, we describe some of the challenges of presenting visual summaries for web search results.

2.6.1 Time required to displaying visual summaries

It was reported that Google tested presenting thumbnails beside traditional search results for 24 hours, then immediately stopped, as results showed a delay in user's responses and hence a decreased number of hits [Hearst, 2009]. However, no details were provided on the effectiveness of thumbnails. As a commercial search engine, Google paid more attention to number of hits, for monetary reasons. The decline in user response in this experiment can be resolved by many variables such as improving the scripts that generate thumbnails and internet bandwidth (speed). However, internet speed is an economic more than an availability issue, and has improved dramatically in recent years [Coffman and Odlyzko, 1998]. In addition, visual summaries can provide the quick gist of an item fast view where 110 milliseconds are enough to get the gist of an image [Woodruff et al., 2001]. This is a good indication that visual summaries may help users to spend less effort on searching. This assumption will be investigated in this thesis. Therefore, researchers should not stop studying the effectiveness of visual summaries merely due to internet speed. Studying the impact of visual summaries not only provides rich information about user searching behaviour, but also

helps to develop approaches for visual summaries.

2.6.2 Visual summary size

The size of visual summaries plays a primary role in their effectiveness, as a small size makes it hard to recognise their features. One of the main issues with thumbnail is that users cannot read textual content due to the small size. Kassten [2002] conducted a study to evaluate the impact of thumbnail size in web browser history. Comparing between the presentation of a title only and a large thumbnail along with a page title, the study showed that users were able to recognise web pages more accurately when larger thumbnails were presented. In addition, results show that users focused on recognising colour and layout when small thumbnails were presented. Results also show that 86% of the time users like to have thumbnails in the browser history. Search result presentation is a significant factor determining the size of visual summaries, where traditional search results present ten items per page. Thus the size of visual summaries should be fitted to the overall structure and limited space should be provided for each item.

However, some existing search engines such as Google use dynamic ways to present their visual summaries for each item of a large size. As the mouse moves over the textual items, a visual summary appears next to the related document surrogate. In our designed interfaces in this thesis, we used a static method to present the visual summaries with size of 200×150 for each item, where selecting the appropriate size was based on the literature review. More details are presented in Chapter 5.

One more issue with visual summaries is the difficulty of generating such summaries for

web pages with a bunch of text, particularly those that do not have images in their content. Visual summaries such as thumbnails will look similar, especially for small thumbnails. However, this issue might be solved by developing an approach for visual summaries that takes this issue into account. In our thesis, we developed a new approach for visual summaries called "Visual tag", where text-only web pages can be easily recognised. (More details about this approach are given in Chapter 5.)

2.7 Summary

In this chapter, we described various approaches for creating visual summaries and studies conducted to evaluate their effects on user searching behaviour. Effective representation of web search results can play an important role in facilitating information seeking. User search performance may be improved by presenting additional visual summaries as part of a web search interface. More investigation is required to evaluate the effectiveness of additional visual summaries. A few studies investigate the effectiveness of visual summaries on user seeking behaviour and user queries, but user searching behaviour was analysed in lesser detail.

In this thesis, we investigate user seeking behaviour in depth to gain a richer understanding of the impact of additional visual summaries on user seeking behaviour. A new approach to visual summary is evaluated in addition to other existing approaches. We investigate the effectiveness of different approaches (including our own) by comparing them with text-only interfaces. Our user studies differ from other research by using real search engine results and designing web search interfaces to simulate a real environment similar to the one that user

may encounter in real life. We design a task scenario that simulates a real task, where a user has a list of search results to select potential relevant answers; in comparison, previous studies focus more on task completion time and error rates. We use advanced techniques (eye tracking metrics) to measure user performance and evaluate user interaction with results screen and effort expended. Finally, we also investigate the impact of topic types on users' information seeking behaviour and the effectiveness of the additional visual summaries.

Chapter 3

The use of eye tracking in information retrieval evaluation

Web search engines are a key enabling technology to support users in finding useful information on the World Wide Web, and the search interface is an important component of these engines. The organisation and presentation of search results is a principal part of the search interface and can have a substantial impact on the ability of users to find their desired information. The evaluation of web search interfaces is therefore essential to present effective information, and eye tracking is a promising technique with the ability to provide rich information for this purpose.

An eye tracker is a device that calculates the exact point of the user's gaze using a geometrical model. It also captures detailed information on timing and click events. In this chapter, we address the first research question in this thesis: How can an eye tracker be used to understand user behaviour when interacting with textual and visual summaries of search

results? This question was consequently investigated, and at the end of each user study, the question was revisited and recommendations were applied on the next user study. Thus, in this chapter we summarise our understanding of the complexities of employing eye tracking and their solutions for the use of eye tracking on web search interface.

In this chapter, a short history of eye tracking is provided in Section 3.1, and in Section 3.2 we describe eye movements and cognitive processes. Section 3.3 describes the use of eye tracking in combination with other measures. We discuss standard and suggested eye movement metrics in Section 3.4, and in Section 3.5 we assess the quality of eye tracking. One of the major issues associated with eye tracking is the eye movement filter, discussed in Section 3.6. In Section 3.7 we describe our approach to using eye tracking for the evaluation of web search interfaces. Section 3.8 discusses the limitation of eye tracking and in Section 3.9 a summary of this chapter is presented.

3.1 History of the eye tracker

The first eye trackers were built in the late 1800s and since that time a variety of different techniques have been applied to collect eye movements [Horrey and Wickens, 2007]. In 1901, Dodge and Cline applied the first method to collect eye movements by reflecting an external light source from the fovea (the center of the macula region, also known as the retina, in the eye) [Richardson and Spivey, 2004]. Since then, many different techniques have been introduced, such as electroencephalography, where electrodes are mounted on the skin around the eyes to capture the eyeball's musculature moves. This allows the observation of eye movements over a large dynamic range; however it can only capture horizontal eye

movements. Other techniques used in the early twentieth century consist of corneal reflection, limbus, pupil, eyelid tracking and the scleral contact lens [Robinson, 1963; Young and Sheena, 1975]. Most of the techniques introduced before the 1970s are invasive, requiring researchers to directly manipulate participants' eyes. However, more recent techniques are non-intrusive, with video images of the eye being used to locate where the user is looking on the screen. This is made possible by the use of infra-red light, which captures and reflects the eye images [Duchowski, 2007]. Nowadays, tracking eye movements can be captured with free head movements by tracking pupil brightness and corneal reflection [Jacob and Karn, 2003; Duchowski, 2007]. Eye tracking is therefore easily applied in human interaction studies, so that the number of studies using eye tracking is dramatically increasing. For instance, in 1987 the first paper appeared in the Conference on Human Factors in Computing Systems using eye tracking by Ware and Mikaelian [1987], and in 2000 in the same conference more than six papers were presented [Jacob and Karn, 2003].

In our user studies, data was collected using a Tobii T60 eye tracker[1]. This non-invasive device calculates the exact point of a user's gaze using a geometrical model. The T60 eye tracker comprises of a wide screen 17 inch monitor with a high resolution. Participants need only look at a computer screen and use a keyboard and mouse to answer the experiment tasks.

3.1.1 Current use of eye tracking

Eye trackers have been widely used in a range of different research fields, including disability studies [Betke et al., 2002; Sears and Young, 2002; Hornof and Cavender, 2005], market-

[1] www.tobii.com

ing [Wedel and Pieters, Fall 2000; Maughan et al., 2007], psychology [Allopenna et al., 1998; Johnson et al., 2003], and applied human factors studies. These applied human factors studies include surgical skills assessment [Richstone et al., 2010], flight management systems [Hanson, 2004; Frische et al., 2011], and driving skills studies [Cohen, 1981; Hughes and Cole, 1986; Martens and Fox, 2007]. However, eye trackers are most intensely used for usability research (evaluating interrelated design) and in the human computer interface (HCI) field. Poole and Ball [2005] classify the use of eye tracking in HCI and usability research into two groups: studies concerning user seeking behaviour and studies focussing on the features of websites related to effective usability.

3.2 Eye movements and cognitive processes

The process of making a decision to find a relevant answer for a web search task involves learning, problem solving, memory and comprehension. These mental processes are called cognitive processes or information processing. Information retrieval studies have demonstrated significant correlations between some cognitive abilities and their indicative factors, such as perceptual speed and spatial scanning. Perceptual speed is defined as speed in finding or comparing figures or symbols, or completing other basic visual perception tasks [Ekstrom et al., 1976]. A study by Allen [1992] finds that perceptual speed can be measured in information retrieval studies by identifying how quickly people scan web content and the speed at which they make judgements. Spatial scanning is defined as the speed at which a viewer explores a complex or broad spatial field [Ekstrom et al., 1976]. Many studies find that task completion time and the ability to select relevant and non-relevant answers are indicators

of spatial scanning and logical reasoning [Vicente et al., 1987; Vicente and Williges, 1988; Campagnoni and Ehrlich, 1989; Jennings et al., 1991; Seagull and Walker, 1992]. Furthermore, studies show that eye tracking can provide significant data about cognitive processes, such as reading, visual searching, and scene perception [Rayner, 1998]. Rayner and Castelhano [2007] found that users spend a longer duration on oral reading and scene perception than on silent reading, while the typical fixation in visual searches is smaller than in other aspects of cognitive processes. Additionally, the type of materials, the difficulty of texts and the goal of the user's reading all influence fixations [Rayner and Pollatsek, 1994; Rayner, 1998].

3.3 Using eye tracking in combination with other measures

Eye tracking can provide stronger qualitative outcomes by working in tandem with other qualitative measures such as think-aloud [Granka and Rodden, 2006] and click through-data [Joachims et al., 2005]. Specifically, using these additional measurements along with eye tracking can improve the quality of behavioural assumptions in understanding the correlation between dependent parameters and user seeking behaviours. Several other techniques that can be used simultaneously with eye tracking are questionnaires (which can be administered pre-, during or post-session), think-aloud and click-through data.

3.3.1 Using eye tracking for enhanced analysis of click behaviour

Eye tracking can capture user strategies for browsing web pages and user attention time on a single item or even a single word. Making use of these findings, some recent studies have

employed eye tracking to investigate the limitation of other measures.

Joachims et al. [2005] investigated the effectiveness of clicks to evaluate web search results, using eye tracking to collect the required data for their analysis. Their study involved two phases: in the first phase, participants were asked to answer ten web search tasks (five informational and five navigational) using Google search engine. In phase two, three manipulated sets of search results were designed for the same ten tasks: (1) using the original ranking list retrieved by Google, (2) using the same original ranking list, but switching the order of the top two items, and (3) using the reversed order of the original ranking list. The analysis of the eye tracking data shows that users were influenced by the ranking order of the search results and consequently often clicked the top two items. In addition, users viewed the search results from the top to the bottom, and paid more attention to items that were above the clicked answer than items that were below on the ranking list. The authors conclude that click-through data is informative, but biased by position.

3.3.2 Using eye tracking in combination with think-aloud

Terai et al. [2008] evaluate user seeking behaviour through the use of informational and transactional search tasks. Participants were asked to think-aloud while answering a given web search task, and their eye movements were captured by eye tracking. Whilst it is true that using think-aloud alongside eye tracking helps us to avoid making assumptions about user seeking behaviour, some are likely to object to this practice on the grounds that user seeking behaviour is (conversely) influenced by think-aloud. For instance, Stephen et al. [2012] investigated the effect of think-aloud on user performance by conducting a user

study involving two groups of participants. One group was asked to practice "think-aloud" as they completed a number of tasks, whilst the second group was instructed to answer the tasks without thinking aloud. Results show that thinking aloud had a significant effect on user performance, as the time to first fixation was significantly longer, and the fixation itself was shorter in duration for the group employing think-aloud. Therefore, in this thesis we did not include think-aloud in our methodology.

3.4 Definition of commonly used eye movement measures in IR

Eye tracking provides rich information on user searching behaviour with the use of two main measurements: fixation and saccades. Those two main data types can be used in a variety of ways to analyse particular aspects of user search behaviour. Ehmke and Wilson [2007] reviewed some of these metrics (fixation, saccade, scan-path and gaze) and the usability problem or cognitive process involved in each. Poole and Ball [2005] provided practical guidance on the use of some techniques, while Moacdieh and Sarter [2012] classified eye tracking metrics to three categories: quantitative, qualitative and the combination of both. Jacob and Karn [2003] reviewed the user usability studies that use eye tracking and the metrics used by those studies.

In this section of the thesis, we reviewed potential eye movement metrics that can be used in information retrieval studies relating to web search results.

3.4.1 Area of interest

An area of interest (AOI) is a region or element of a screen, which allows researchers to measure the amount of attention that a user spends on a precisely defined exact object. Several components of the eye's movement can be measured using AOIs, such as the first view of an AOI, or the frequency of views.

Furthermore, AOIs can compare the user's gaze on each separate visit to the same AOI: for example, comparing the time that a user spent reading a particular text item before and after selecting an answer for a given search task. In addition, AOIs can be measured by several fixation parameters such as total number of fixations, fixation duration, first fixation, fixation order and frequency. The basic AOI events are hits, dwells (total viewing time) and transitions (eye movements from one AOI to another) [Holmqvist et al., 2011].

One of the significant benefits that AOIs provide in user studies is the identification of details about user strategies for browsing a webpage. These details enable us to draw a diagram modelling user seeking behaviour, allowing us to represent the details of how users interact with webpage elements.

3.4.2 Fixation and saccades

Eye movement measures evaluate eye location and the attention spent on a single portion or an entire region of the screen. The main data that eye tracking captures to represent the location of a user's gaze on the screen falls into two types: *fixations* (a series of gaze points maintained in a constant direction whilst on individual is observing a stationary target); and *saccades* (quick eye movements between fixations). A fixation is the duration of time spent

viewing a particular target located on the foveal range of human eye, while saccades connect between fixations. Fixations can transmit visual signals to the brain, but saccades cannot do so [Ellis, 2009].

Eye movement can be analysed through the use of many quantitative measures. The majority of these quantitative measures are reviewed in this chapter, and Goldberg and Kotvel [1999] Poole and Ball [2005] also provide a review. For example, eye fixations, which provide deeper insight into users' attention, can be measured in multiple ways: time to first fixation, duration and total number of fixations. These numerical measures play a significant role in interpreting user behaviour. These measures provide different explanations of behaviour: for example, higher fixation frequency on a particular object indicates a noticeable or important object, while a long duration of fixation suggests a difficulty to process information. Goldberg and Kotvel [1999] investigate these measures in depth and conclude by identifying three categories of measures that are potential indications of user seeking behaviour: global search, local search and processing.

Fixation-related metrics are widely used throughout user studies. For the purpose of analysing the fixation-related metrics that are reported in the literature, we can divide them into three categories: those that measure (1) difficulty in extracting information, (2) interface efficiency, and (3) user viewing behaviour (i.e. , whether the user is processing or searching). Difficulty in extracting information can be measured by the time of arrival (time to first fixation) and fixation duration and length [Joachims et al., 2005; Huang and Gordon, 2011].

Saccade/fixation rates provide an indication of whether a user is processing (fixation), or searching (saccades). More saccades indicate more searching behaviour [Goldberg and

Kotval, 1999; Liversedge and Findlay, 2000; Poole et al., 2005]. Goldberg and Kotval [1999] found that good interfaces produce significantly lower numbers of fixations, but fixation duration and fixation/saccade ratio were not significantly variable in comparison with poor interfaces. Efficiency can be measured by fixation spatial density, where the closer fixations suggest that a user is finding interesting information [Cowen et al., 2002].

In addition, these metrics can be employed to produce more metrics such as the overall number of fixations [Goldberg and Kotval, 1999], total number of fixation on each AOI [Poole et al., 2005; Cutrell and Guan, 2007], total/mean fixation time [Cutrell and Guan, 2007; Kammerer and Gerjets, 2010; 2012] and number of users fixated on a particular AOI [Albert, 2002; Granka et al., 2004; Joachims et al., 2005]. In addition, dwell (the total gaze spent on a particular AOI) is employed in many studies [Jacob and Karn, 2003; Oertel and Hein, 2003] for purposes such as measuring the processing time or difficulty of a given task.

3.4.3 Heat maps and gaze plots

Eye tracking data can graphically represent a user's attention on a screen using various techniques, such as a *heat map* (a graph that displays gaze data in a matrix, using colours to represent attention on screen) and a *gaze plot* (a graph that represents the sequence and order of eye movements and their duration). An example of a gaze plot and heat map for Google search results page are shown in Figures 3.1 and 3.2. Heat maps allow us to recognise the most viewed areas on screen, providing an indication of important or attractive objects. Gaze plotting provides detailed information about user searching behaviour such as first fixations, viewing duration and exact viewed locations.

Figure 3.1: Gaze Plot: (A) Area of interest. (B) Fixations. (C) Saccade. (D) First fixation.

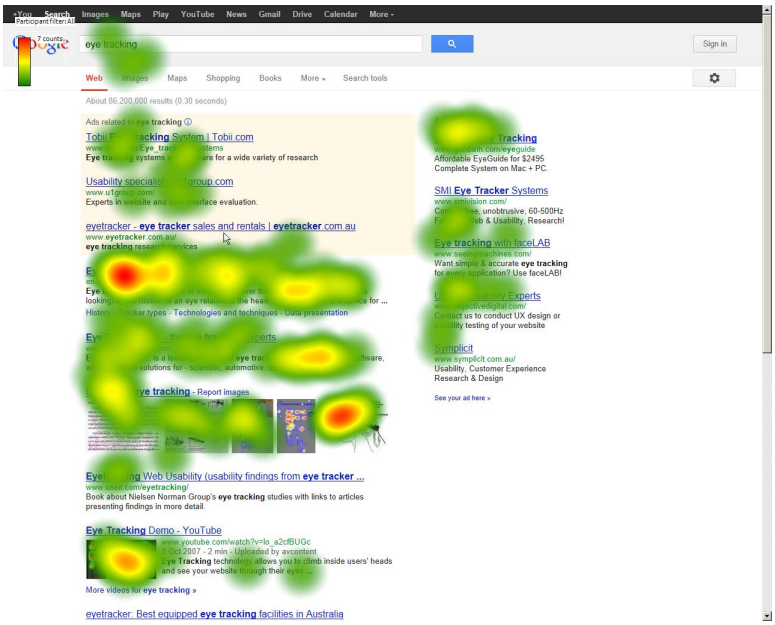

Figure 3.2: Heat maps visualise user gaze data by using colours. Dark colours represent a large concentration of gaze and lighter colours represent a lower concentration of gaze.

Nielsen [2010] and Soussan et al. [2011] describe how users view web pages according to the outcomes of heat maps generated by an eye tracker. Their results show that users view the left and top sides of web pages more than other areas. Shretha and Kelsi [2007] investigate how users browse the results of e-commerce websites by observing the visual scan paths of users. Their analysis focuses on describing the outcomes of hotspots and gaze plots that illustrate users' visual attention. They found that visual attention is uniform during the browsing of picture results, but follows an "F-shaped pattern" for browsing and searching text on web content. In other words, users spend more attention on top ranking text items and view them more carefully, while spending less time and a quicker view on low ranking items and viewing them more rapidly. Moe [2007] uses eye tracking data to qualitatively identify features of implicit relevance feedback. Results show that thorough reading (such as careful reading) is one of the most useful methods to find relevant information.

Heat maps allow us to analyse eye tracking data in a purely qualitative format. For instance, qualitative analysis of eye movements can help to identify the most attractive viewed elements, the start point, dominant seeking behaviour and missed elements. Whilst we can evaluate the effectiveness of an interface by identifying the time required for given tasks, qualitative eye movement measures can provide the reasons behind the inefficacy of a specific portion of the interface. An eye tracker allows us to conduct qualitative screening studies (studies that use heatmaps to analyse user behaviour), which help us to gain a richer understanding of viewing patterns. However, Bojko [2009] argues that heat maps should not be used on their own without quantitative measures. She investigates the use of heat maps in describing and interpreting user behaviour and concludes by giving some guidelines on how

to use heat maps more effectively. For instance, it is better to use fixation instead of raw data to generate heatmaps, and within a study, a definition of fixation, such as minimum fixation duration, should be consistent during the analysis and visualisation of the data.

In our studies, we used heat-maps and gaze plots to explore how users browse search results, which helped us to implement a diagram to observe user behaviour and attention. More details about this technique are explained in Chapter 6. In addition, we use gaze plots to evaluate the quality of calibration, as explained later in this chapter.

3.4.4 Scan-paths

An eye tracker gathers two main measurements to represent the point of a user's gaze on the screen: fixations and saccades. A *scan-path* is a completed observed path of eye movement sequences (fixations and saccades) across a screen. Scan-paths can provide valuable insights into information seeking behaviour and cognitive styles, including users' mental effort. Noton and Stark [1971] proposed the scan-path examining how user viewing images. Results show that the amplitude of saccades in images is similar to that in reading text, and fixations on viewing images are uniform. Josephson and Holmes [2002] investigated the scan-paths of eight subjects for three different web page domains, and found that users have "habitually preferred scan-paths". In other words, individuals favour certain scan-paths, but factors such as type of web page can influence their choice. Investigating the features of scan-paths can provide us with good indications about user interaction with informative components, and the effectiveness of such components. Poole and Ball [2005], and Goldberg and Helfman [2010b] discuss several metrics to quantitatively analyse scan-paths and their accompanying methods

of interpretation, such as scan-path duration and length.

Scan-path length: Scan-paths consist of gaze samples, each of which has coordinates on the screen, so that counting the distances between gaze samples will provide us with the scan-path length. Scan-path length is measured in pixels.

Scan-path duration: The total number of gaze points (n) multiplied by gaze duration (0.017 seconds) provide scan-path duration, while for sequences of saccades and fixations, scan-path duration is the sum of the durations of both fixations and saccades [Goldberg and Kotval, 1999].

Goldberg and Kotval [1999] conducted a study to compare between good and poor interfaces by analysing eye movements of twelve users. The results show that no significant differences were found in scan-path duration between good and poor interfaces; however, poor interfaces produce significantly longer scan-paths. Goldberg and Helfman [2010b] propose a technique to automatically identify similar scan-paths in order to study user browsing strategies. Their study provides an opportunity to analyse scan-paths qualitatively, particularly for studies with many participants or factors [Goldberg and Helfman, 2010a]. We therefore, we used scan-paths to analyse our collected data, as will described later in this chapter.

3.4.5 Transition rate (re-viewing)

Transition rate is a useful metric particularly for free viewing, where saccade direction is examined [Ponsoda et al., 1995]. AOIs allow us to observe the transition rate between informative components (frequency of eye movements between AOIs), where a higher rate of transition indicates less efficient management between those components [Jacob and Karn,

2003]. A forward and backward gaze movement indicates that the user is not sure about their search or about the presented information [Goldberg and Kotval, 1999].

In our user studies, we improve the transition rate definition in order to produce the re-viewing percentage for items (such as the number of times a user re-views an item). In our designed interfaces, five search result items was presented for each search topic. AOIs were used to collect the total amount of times items were viewed by users (*Total viewed*), and the number of uniquely viewed items out of the possible five search results was also collected for each session (*Uniquely viewed*). The following formula was then used to produce a re-viewing percentage:

$$\text{Re-viewing} = \frac{Total\ viewed - Uniquely\ viewed}{Total\ viewed}$$

This metric examines users' mental search behaviour and measures the effectiveness of the presented search results, applying the same suggestions of transition rate, higher re-viewing percentages indicate searching and lower effectiveness [Goldberg and Kotval, 1999; Jacob and Karn, 2003]. This proposed metric is applied in Chapter 6.

3.4.6 Other eye tracking metrics

Another metric, *spatial density*, measures user searching, where extensive gaze points indicate inefficient searching whilst small samples of gaze points reflect an efficient search. Goldberg and Kotval [1999] analysed the scanned area using the Convex hull; their results show that users scanned significantly more areas using the poor interface rather than the good interface, while the poor interface produced significantly larger spatial density than the good interface.

Pupil dilation and blink rate can also be captured by eye tracking, and studies show these

metrics can be used to indicate cognitive workload [Bailey et al., 2007; King, 2009; Bruneau et al., 2002; Riding and Rayner, 1998]. A cognitive workload can be identified by a lower blink rate; in contrast, a higher blink rate is a sign of stress and overwork [Poole and Ball, 2005]. Unfortunately, pupil size and blinking can be influenced by many factors, such as the light source used, and therefore a few studies apply these metrics in eye tracking research. Jacob and Karn [2003] argue that this issue can be solved by using a specific type of hardware device, IBM Blue Eyes; however they admit this will not address the issue completely. We therefore do not consider using pupil dilation and blink rate in this thesis.

3.5 Quality of eye tracking

Quality of eye tracking is influenced not only by mechanical setup and pupil detection, but also by other factors, which may include [Drewes, 2010]: (1) accuracy (influenced by the eye's anatomy, including muscles and nerve sensors such as heat and cold), (2) low-level filtering and (3) the type of application software attached to eye tracking for the evaluation and visualization of the collected data. In addition, the quality of eye tracking can be influenced by time resolution (frame rate of a video), latency (delay in capturing the data) and robustness (impact of light and users' glasses or contact lenses).

3.5.1 Types of eye movements

Jacob and Karn [2003] classified two categories of eye movements based on the environment of the study: natural and unnatural. The most common type examined in studies is natural eye movement, where users freely browse web documents, with no particular instructions

avoiding their viewing. Unnatural eye movements occurs where users are instructed to follow a specific way of viewing a document.

3.5.2 Calibration

Calibration must take place before participants start the experiment, because characteristics of the eye (such as pupil size and nerve sensors) differ from one participant to an other. Calibration is used to detect the participant's gaze point and to create a personal profile for each participant's eye features. The calibration is conducted by moving a dot to different locations on the screen (calibration points), and the participant is instructed to follow that dot. The number of calibration points can be increased.

Unfortunately, eye tracking systems do not provide any criteria to check the quality of calibration. Some eye tracking manufacturer software such as "Tobii Studio"[2] provides a percentage of recording quality, which is based on the total captured gaze time out of total time user spend answering a given task. This percentage provides a good indication of the amount of missing gaze that eye tracking does not manage to capture, most probably where users did not look at the screen. However, this percentage does not affect the quality of calibration, and can be only examined manually by checking the gaze plot and target's location. This can be done for a small amount of sessions but is not sufficient for a large number of sessions, and particularly not for web search interfaces, where targets are the informative components. For instance, in our user studies, targets are visual summaries and textual summaries (a web page title, a short text extract from the source document, and the URL). In addition, on a web search interface, users engage in diverse behaviours such as

[2] www.tobii.com

reading and skimming or searching and processing. These different behaviours make it hard to identify the quality of calibration. The proposed techniques that have been developed to solve bad calibration require manipulating the gaze data positions using average gaze point, locations to identify the fixation location [Hornof and Halverson, 2002].

One study recommends recalibrating at the middle of the study to make sure of good calibration [Aaltonen et al., 1998]. Others suggest to re-calibrate at any time during the experiment when it is necessary [Pollatsek et al., 1990], however, these suggestions are uncomfortable, interrupt users and can mislead users about the actual goal of the study. Some studies calibrate at the beginning of each session and verify the quality of calibration before each trial [Abrams and Jonides, 1988; Abrams et al., 1989]: in these studies, if the eye was not on the right fixated point by 1 °, then participant was asked to repeat the trial. However, in experiments such as our user studies, participants are not instructed to look at specific paths or locations; they are only instructed to look at the screen. Direction degree cannot be applied in studies where users are freely looking at the screen and targets have different dimensions: position and size.

3.6 Eye movement filters

Identifying fixations and saccades in eye tracking protocols is an essential part of user studies, and can significantly impact upon the overall indication of the results [Salvucci and Goldberg, 2000]. Fixation filters cluster the gaze data points to meaningful fixations. Many algorithms are proposed to define fixations, such as the Adaptive algorithm [Smeets and Hooge, 2003], Kalman filter identification (I-KF) [Sauter et al., 1991], and algorithms developed by eye

tracking manufacturers such as Tobii's fixation algorithms (ClearView and Tobii Fixation Filter) [Larsson, 2010]. To group gaze data into meaningful fixations or events, various aspects are taken into account such as the distance between two gaze points, gaze position, data sequence, radius and duration of fixation [Komogortsev et al., 2010].

Several studies investigate fixation identification (the minimum required time for a fixation); some have found that the minimum required time for fixation is 100 to 150 milliseconds, and the typical average length is longer than 150 milliseconds [Kowler, 1990; Yarbus et al., 1967], while others suggest the minimum fixation is 200 milliseconds [Fischer and Ramsperger, 1984; Jacob and Karn, 2003]. A number of researchers defined the minimum fixation duration as 100 ms in their user studies [Karsh and Breitenbach, 1983; Goldberg and Kotval, 1999; Salvucci and Goldberg, 2000; Holm and Mäntylä, 2007; Goldberg and Helfman, 2010b]. Fixation identifications can be influenced by several factors: Mackworth [1967] found that display densities exert an influence, while Friedman and Liebelt [1981] noted that fixation durations are longer on objects that require difficult processing. Furthermore, Salvucci and Goldberg [2000] argue that fixation identification is still subjective, and they contend that fixation duration on visual processing is different to that in cognitive processing. In summary, algorithms of fixation identification should take these aspects into account to easily compare between previous studies.

Additionally, some studies use different filters according to the subject material and focus of the study. Because of the sensitivity of identifying fixations and saccades, several studies develop techniques to select an appropriate filter to get more accurate data [Falkmer et al., 2008; Juhola et al., 1985; Sauter et al., 1991]. Furthermore, Marcus and Kenneth [2010]

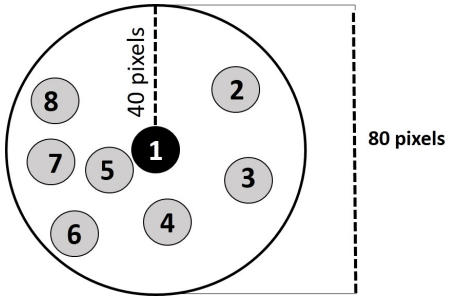

Figure 3.3: A fixation must include at least six gaze points, and the distance between them must be less or equal to 40 pixels.

develop an algorithm to identify eye movement events, such as reading and scene perception.

In our user studies, fixation was defined as a series of six gaze points falling within a 40 pixels radius [Goldberg and Kotval, 1999] (see Figure 3.3). In addition, other studies agreed that the minimum of fixation duration is 100 ms [Karsh and Breitenbach, 1983; Salvucci and Goldberg, 2000; Holm and Mäntylä, 2007; Goldberg and Helfman, 2010b]. In our studies, users were asked to predict the relevant answers: this process involves variables of cognitive workload and behaviour such as searching, reading, scanning and selection. These different aspects can influence fixation durations, therefore, we did not use fixations and saccades metrics in our analysis (more details are discussed in the next section).

3.7 Our approach to using eye tracking for the evaluation of web search interfaces

Eye tracking cannot be perfect [Aaltonen et al., 1998]. Each field has its own character and variables that should be taken into account when conducting a user study. Using eye tracking in web search interfaces is not the same as using eye tracking to evaluate websites or software interfaces. The searching process that users follow has specific and distinct aspects.

Search result items should be controlled to produce more precise results. We used eye tracking to evaluate existing web search interfaces, and results show that a number of factors need to be controlled in order to get more precise results. We therefore designed a web search interface to evaluate the effectiveness of presenting additional visual summaries. One significant factor affecting the use of eye tracking for the evaluation of web search interfaces is the space left between informative components. Enough space must be left for researchers to easily distinguish user gaze data in the analysis stage. Another recommendation is to use reasonable font and image size, so that users do not have to spend longer fixations to recognise objects of smaller size. Scrolling pages can be annoying in the analysis of gaze data where users may devote some attention to scrolling through results pages; hence, in our interfaces, we present a reasonable number of result items that fit on one non-scrolling web page.

In addition to the above recommendations, we used several techniques to improve our findings and user gaze data. The following techniques were consequently applied in our user studies.

3.7.1 Using questionnaires

Questionnaires can provide good feedback, improving findings and making assumptions and interpretations of predictable patterns of user searching behaviour more robust. Questionnaires can take place at any stage during the experiment, depending on the feedback and search task. Studies show that eye movements, fixation and saccades can be influenced by familiarity of information [Greene and Rayner, 2001; Williams and Morris, 2004], and in

addition, as discussed earlier, researchers found that user difficulties produce longer fixations and intensive gaze points. We therefore strongly suggest asking users about the difficulty and familiarity of information on web search tasks when eye tracking is used. One means of obtaining feedback on difficulty is to compare between fixation durations and user responses. In our user studies, user feedback on the perceived difficulty and familiarity of each search was collected: participants were asked two questions at the end of each session. The first investigated their familiarity with the topic, and the second concerned the difficulty they experienced in finding relevant answers.

3.7.2 Adaptation of gaze direction

Users move their eyes according to the content displayed on the screen. The first fixations are strongly correlated with previous direct gaze [Palanica and Itier, 2011], and gaze perception is influenced by the adaptation of gaze direction [Jenkins et al., 2006; Schweinberger et al., 2007; Kloth and Schweinberger, 2008]. Therefore, viewing one media and then another during the capturing of user eye movements can influence the gaze' starting point on the second media. For example, if a user was reading a task instruction located on the bottom of the screen, and then immediately a search results page was displayed, their eye movements will show that the initial gaze point on the search results page starts at the bottom. Thus it is recommended to display a black or white screen before showing a search results page. In our study, we display a black screen for three seconds before displaying search results.

3.7.3 Evaluating the quality of calibration

We developed a technique to help evaluate the quality of calibration for our user studies. Determining the quality of calibration is significant, particularly for web search interfaces where information is presented in a strict space. Incorrect calibration, whether systematic or variable (depending on error rate), can provide completely wrong assumptions and results. For instance in the use of textual surrogates (combination of titles, snippets and URLs), when systematic errors occurred while users were looking at titles, gaze points were instead located on textual snippets. Researchers should therefore be aware of the importance of evaluating the quality of calibration.

The average number of sessions in our user studies is 250 sessions, and hence it is difficult to check individual session calibration manually. We used a technique to provide us with a signal of how good the calibration was for each participant.

Our technique involved presenting one page displaying only one text line at the beginning of the experiment, and another page displaying only one text line at the end of the experiment. The idea was that when users read the text line at the beginning and the end of the study, we could receive an overall signal of the quality of the calibration. Some manufacturer software (such as Tobii Studio) enables researchers to generate gaze plots for all media with only one click: thus it is easy to check the quality of calibration for a page with only one text line by using a gaze plot. This technique not only provides a hint of the quality of calibration, but can also provide an indication of whether there is a variable error or systematic error. Figure 3.4 shows an example of the described web page.

In addition, this technique can provide a good suggestion of how participants were acting

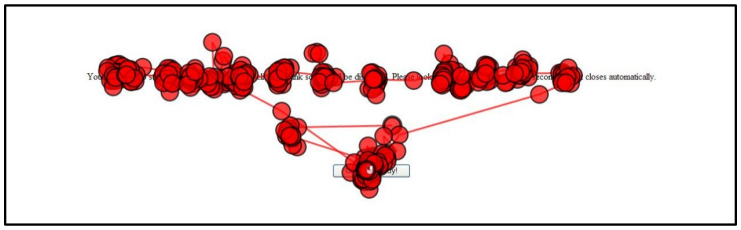

Figure 3.4: Example of the page with one text line used to check quality of calibration.

during the study, where if the text is not read then participants can be biased, and a further investigation takes place. We eliminated participant data with systematic errors. If the collected gaze data had a variable error, then we manually checked the proportion of error. If the variable error was a reasonable proportion (determined by manually viewing the gaze plot) then we kept the trial.

3.7.4 Optimising collected eye movements

Eye tracking may lose the capacity to coordinate a user's attention for a while due to many reasons, the most common of which are head movements, the loose capture of pupils/corneas and the interruption of viewing for typing or clicking a mouse. A cut or a missing patch in eye movement data can mislead researchers. Only six gaze points ($0.017s \times 6$) produce a fixation [Karsh and Breitenbach, 1983; Goldberg and Kotval, 1999; Salvucci and Goldberg, 2000; Holm and Mäntylä, 2007]. Gaze point errors can be classified into two types: "systematic" errors (where the gazepoint data location is not correct, Figure 3.5 (B)) or "variable" errors (where the distance between gazepoint locations appears larger than it is, Figure 3.5 (C)) [Chapanis, 1951]. Only a limited selection of papers discuss solutions for gaze point errors, yet this can significantly impact on the overall outcomes of the study [Hornof and Halverson, 2002].

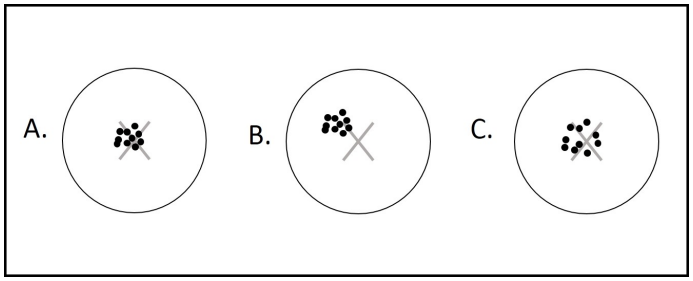

Figure 3.5: Eye tracking errors: (A) Good captured gaze points (B) Systematic error (C) Variable error.

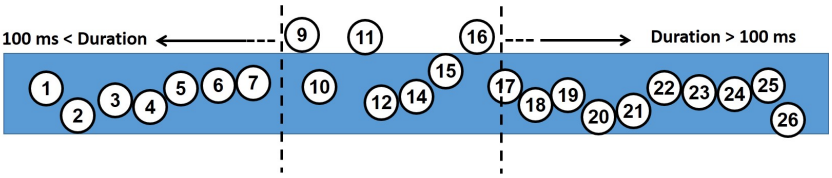

Figure 3.6: An example showing an AOI (box) and gaze points (1-26). According to our approach, the gaze points (9, 11 and 16) are assigned to this AOI.

To address this problem, we develop a technique to improve the quality of collected eye tracking data. An algorithm was formulated to optimise the collected gaze data before analysing user behaviour and web search interfaces in the experiment. The goal of the algorithm is to partly reduce the impact of variable error. The idea is that variable errors cause an increase in the distances between gaze point data, which may move some gaze points outside the AOI coordinate. For instance, some gaze points can be just located out the AOI border (see Figure 3.6). When a fixation is 100 ms (six gaze points), losing one gaze point out of the count in an AOI can eliminate a fixation. This can influence analysis of the results: for example, if one fixation is occurred on a particular AOI, such as a visual summary, then analysis will take into account that the user viewed that item to answer a given task.

In our algorithm, firstly, the gaze point data was assigned according to its position in one of the AOIs or white-space. Fixation was defined as a series of six gaze points falling

within a 40 pixels radius. Secondly, up to three white-space gaze points were re-assigned to an AOI if a series of six gaze points (at least) was assigned to the same AOI before and after the re-assigned gaze points, see Figure 3.6. Therefore, if one, two or three gaze points were identified in white-space occurring between two series of six gaze points identified for same AOI, then we change the points identified on white-space to that AOI. We manually checked a sample of the manipulated white-space gaze points positions, and results show that they ocurred just next to the AOI's borders.

3.7.5 Complexity of employing fixations and saccade metrics

Camilli et al. [2008] found that increasing the threshold (minimum duration) results in missing some fixations, while decreasing the threshold produces false fixations. Studies recommended that a threshold be set to 100 ms [Manor and Gordon, 2003; Radach, 1998]. Poole and Ball [2005], who review various metrics of eye tracking, agreed with that recommendation, as do other researchers [Karsh and Breitenbach, 1983; Goldberg and Kotval, 1999; Salvucci and Goldberg, 2000; Goldberg and Helfman, 2010b; Holm and Mäntylä, 2007].

Some studies show that 40 ms is enough for users to get the gist of a scene [Grill-Spector et al., 2000; Castelhano and Henderson, 2008], and others suggest that 50-60 ms is enough to read a fixated word [Ishida and Ikeda, 1989; Rayner et al., 2003; 2006]. In addition, research shows that some aspects related to text influence fixation duration, such frequency of words [Inhoff and Rayner, 1986] and familiarity of words (where fixation can be less than 100 ms) [Reichle et al., 1998; 2003]. Moreover, eye movements are influenced by the distinct type of visual behaviour in activities such as skimming. This can take place in

the same manner when recognising images, where familiar images can be recognised in a shorter amount of time than the first time a user sees an image. Our user studies involve both text summaries and visual summaries: therefore, in our analysis, we used metrics that involve gaze points instead of fixations and saccades to analyse user searching behaviours, where measures such as dwell and scan-path help to analyse the gaze data. We did not use fixation and saccade metrics in our analysis because they can be influenced by the variety of user searching behaviour and the features of the presented information (such as types and variety of summaries and familiarities). For instance, our user studies involve popular visual summaries such as the *thumbnail*, and users are more likely to have seen these visual summaries previously; however, they are not likely to have seen a *visual tag*, which is a new approach to visual summaries. Furthermore, selecting the right filter requires further investigation to produce accurate identification of fixations where, for example, most existing algorithms rely on taking the average of gaze point locations to identify the fixation position.

3.7.6 Using scan-paths

In our user studies, scan-paths are calculated for each area of interest (AOI), where each informative component on the tested interface was defined as an AOI. The most common way to define a scan-path is the eye's path from the moment that a user starts browsing the web page until the end of the task. However, in our user studies, the beginning of each scan-path is determined by the moment at which the user starts to look at the defined AOI and ends at the point when the user leaves that AOI. The study of scan-path requires manual analysis of the data instead of computer algorithms [Wetzel et al., 1996; Byrne et al., 1999;

Goldberg et al., 2002; Goldberg and Helfman, 2010b]. Consequently, we develop our own technique to analyse the gaze data (more details are provided in next section).

3.7.7 Our proposed algorithm to processing raw gaze data

The manufacturer of the eye tracking software (Tobii Studio) provides a percentage for recording quality that is based on the total captured participant's gaze out of the total time required to answer the given tasks. Due to interruptions and calibration issues with the eye tracking, we eliminated users with less than 80% capture accuracy. We exported raw data of the high quality recordings to text data where each user data was stored on a text file separately. As each participant was asked to carry out a series of tasks, we identify the beginning of each session. The loading time of the search results page was excluded from the analysis. We cut down the interfaces involved to target areas (informative components) and non-target areas (white-space and search box). The exported text data provides a set of properties for each gaze point in the data. Examples of properties include the coordinate values of both eyes on the screen, and the validity of the gaze data in the tracking of both eyes.

1. Removing noise data.

   This is a normal step in employing filters and algorithms of eye movements, due to the interruptions and difficulty with calibrating the eye tracking whilst users answer the given task. Two criterias were applied in this process. Firstly, any gaze point occurring out the screen range (where the user was not looking at the screen) was removed. Secondly, the validity of each gaze point was checked to determine if eye tracking could capture both eyes or not. These two criteria are correlated with the recording quality,

and as the gaze of our participants was of good quality(greater than 80%), the removal of noise data did not have a significant impact.

2. Identifying gaze points that occur in AOIs.

   Each AOI has coordinates on the screen that allow us to identify any gaze point within its area. The locations where the user' gaze rested were collected by a mask where each informative component's region was bounded (AOI), leaving regions of white-space. If gaze point data did not occur in one of the AOIs, then it was assigned to white-space.

3. Optimising gaze data.

   In our analysis, we split the screen into targets (AOIs) where gaze data was collected. Some gaze points occurred just outside or on the borders of the AOIs due to a variety of factors such as variable error. We therefore developed a technique to optimise the detection process to include these gaze points (no more than three in total) as falling within the AOI, provided that they meet the applied criteria. More details about this techniques are found in Section 3.7.4.

4. Identifying viewed AOIs.

   Fixation was defined as a series of six gaze points falling within a 40 pixel radius [Goldberg and Kotval, 1999]. In our analyses, we state that an AOI has been viewed if at least one fixation occurred on that AOI. The fixation is the signal of stationary gaze. We used this fixation definition to determine whether an AOI was viewed or not: if it was viewed, then we started collecting the gaze points that were spent on that particular AOI.

3.8 Limitations of eye tracking studies

Aside from the relatively high cost of eye tracking hardware, there are some limitations on eye movement measures due to usability issues and difficulties with some of the tracking techniques. Some of these limitations are related to the complex correlation between eye movements and cognitive processing. Other are attributable to eye movement algorithms and the software used by eye tracker manufacturers.

3.8.1 Lack of overall standardisation

Terms for the measures and properties of eye trackers are far from uniform for a variety of reasons. The software of the manufacturer for example, may cause variation in terminology [Holmqvist et al., 2011; Duchowski, 2007]. Another cause of confusion is that many names derive from different areas of research: for instance, in reading, gaze duration is the amount of attention spent on a particular AOI (also referred to as fixation duration), yet in human factors research, this measure is called a dwell or glance [Green, 2002; Horrey and Wickens, 2007]. The similarity in concept between the two measures becomes obscured by the competing terminology, and comparing new studies with previous research becomes a lengthy and time-consuming process − whereas a standardised term would streamline analysis.

3.8.2 Using different detection algorithm parameters

Many research papers show the results of their analyses using eye tracking data, but a small number of papers outline the parameters of the detection algorithm that are used to define the fixations and saccades in their data. Using slightly different values for the parameters of the

detection algorithm can create dramatically different results [Karsh and Breitenbach, 1983]. Salvucci and Goldberg [2000] investigate different ways to identify fixation and saccades, and one interesting outcome of their study is that there is as yet no standard method for identifying fixations and saccades. Therefore, researchers cannot easily compare two studies without first identifying the eye tracking protocols used in detecting fixation and saccades in the data.

3.8.3 Peripheral vision

Eye tracking can only capture the point of gaze − the eye's position on the screen. This does not include peripheral vision (or the larger area that surrounds the specific point at which the eye is looking), which operates at a lower resolution. To our knowledge, based on available research, no study investigating the effects peripheral vision on eye tracking analysis exists. However, generally speaking, users cannot read or view an object to extract information without clearly seeing that object. Kelly and Cool [2002] investigated the difference in user ability in word recognition between central and peripheral vision. They found that a user requires two to four times the amount of time spent in central vision to recognise a frequent word in peripheral vision. Some studies agreed that the minimum attention required by users to understand or view an image is eligible to be captured by an eye tracker [Haber and Hershenson, 1973; Young and Sheena, 1975].

Furthermore, recent studies provide some techniques to distinguish types of reading behaviour such as skimming or carefully reading [Campbell and Maglio, 2001; Reichle, 2000; Frazier and Rayner, 1982; Hyönä et al., 2002]. For instance, Hyönä et al. [2002] studied

users' reading strategies by analysing the eye movements of 48 participants. Results show that reading strategies can be classified into four types: (1) non-selective reviewers (looking back and forward at sentences), (2) topic structure viewing (reading only introducing-topic sentences), (3) fast and (4) slow linear reading. Reichle et al. [1998] developed a model (E-Z Reader) to understand user reading behaviour including the identification of words and visual processing. Campbell and Maglio [2001] developed an algorithm to detect user reading behaviour (reading, skimming and scanning).

3.9 Summary

In this chapter, we discuss the use of eye tracking in research and employ eye tracking in combination with other research methods. In addition, we review potential eye movement metrics for web search interfaces and their interpretations. We also discuss the methodology we use to analyse the gaze data in our studies, and suggest new techniques to optimise the quality of collected gaze data.

Chapter 4

Interaction with novel web search interfaces

Search result organisation and presentation is an important component of a web search system, and can have a substantial impact on users' ability to find useful information. Most interfaces include textual information (including for example the document title, URL, and a short query-based summary of the content). Other interfaces include additional browsing features such as topic clustering, or thumbnails of web pages. In this chapter we describe a study using eye tracking to compare the effectiveness of three publicly available search interfaces for supporting navigational tasks. The three interfaces vary primarily in the proportion of visual versus textual cues that are used to display a search result.

In this chapter, we begin to investigate the second research question in this thesis: Does providing additional visual summaries for the presentation of web search results impact on users' information-searching behaviour and performance? We conducted a user study that

involved carrying out a series of named-page finding tasks using a variety of search interfaces to investigate the following sub-questions:

1. How do different search interface features impact on users' information seeking behaviour, particularly on the time spent on finding desired information?

2. Do users find looking at text more useful than looking at visual representations?

The chapter is organised as follows: our experiment design, including the different search interfaces, users, procedure and topics used, is described in Section 4.1, and in Section 4.2 we describe the features of the search interfaces involved in the study. The results of the experiment are analysed in Section 4.3, and a discussion and summary are given in Section 4.4.

4.1 Experimental framework

The purpose of this study is to investigate whether and how different search interface features impact on users' information seeking behaviour. In particular, this study emphasises the relationship between those features and the time spent on finding the desired information. The study also investigates the relative attention that users pay to different interface components. We consider three interfaces and three navigational search topics for this purpose.

4.1.1 Experimental setup

The participants in our user study were visitors who attended Open Day at our university in August 2009. They were mostly high school students, and all had some interest in computer science. Visitors were provided with a one-page plain language statement outlining the goals of the experiment, the procedure of the study, and what kind of data would be collected.

Based on this information, visitors could choose to participate in the experiment or not. In total, 35 visitors volunteered to participate in the study. Participants were given no training with the selected search interfaces, and were unlikely to have used them before.

Each participant undertook three navigational search tasks using different search interfaces. Information about the visual attention given to the different screen components was collected using a Tobii T60 eye tracker. This device uses the reflection of near-infrared lights in the eyes to enable non-intrusive tracking of gaze position on a computer screen. It also captures detailed information on timing and click events.

4.1.2 Interfaces

For our study, we selected three web search engines that users were unlikely to be familiar with: Carrot2 (C) [1], Middlespot (M) [2], and Nexplore (N) [3].

The interfaces were selected because they represent a variety of features that go beyond the "default" ranked-list style of search results made popular by systems such as Google, Yahoo!, MSN and Bing. Carrot2 is a clustering engine that organises a search result into thematic categories. Middlespot uses both text and thumbnails when presenting a search result, with the space given to the thumbnails being substantially larger than the text area. Nexplore displays similar components to major search engines, but renders them in a predominantly visual fashion, as shown in Figure 4.1.

Carrot2 (C): Carrot2 presents its search results in the traditional way (title, snippet and URL), as shown in Figure 1 (labelled area 2). However, Carrot2 also clusters the results

[1] http://www.carrot2.org
[2] http://www.middlespot.com
[3] http://www.nexplore.com

Figure 4.1: Nexplore interface (`www.nexplore.com`): (1)Suggested or related queries; (2) Thumbnails; (3)Text area, and (4)Sponsored links.

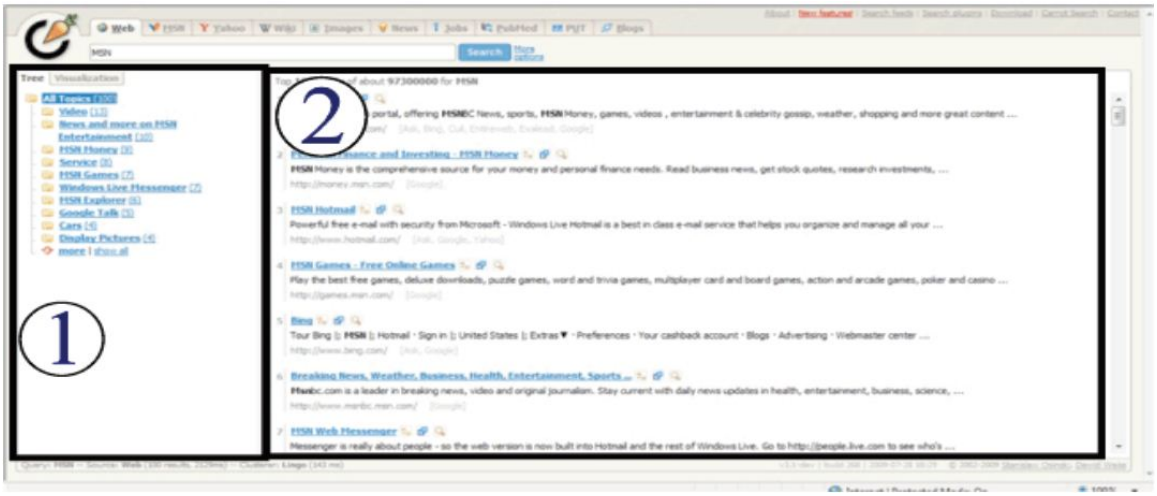

Figure 4.2: Carrot2 interface (`www.carrot2.org`): (1) Text area, and (2)Screenshots of web pages.

Figure 4.3: Middlespot interface (`www.middlespot.com`*) (1) Text area, and (2)Screenshots of web pages.*

and allows the user to browse the clusters in two ways: a hierarchical tree structure (area 1), and a visualisation, where results are presented on a dynamic map for common, popular and other potentially related facts. The visualisation feature is not shown in Figure 4.2 as it was almost never used by our subjects.

Middlespot (M): The Middlespot interface, shown in Figure 4.3, is divided into two areas: a text area (area 1) that presents the title, snippet and URL, and a visual area (area 2) that shows screenshots of web pages. A substantially larger proportion of the screen area is devoted to the visual features (around 70%, while the text area uses less than 20%). When the mouse is moved over a specific screenshot, the corresponding image is enlarged (as shown in area 3 in Figure 4.3). In addition, the corresponding text summary is activated in the left-hand pane, and vice versa, if the mouse hovers over a text summary. This leads to a lot of movement on the screen, as snapshots are enlarged and the text summary list scrolls around to the item that is currently in focus.

Nexplore (N): As shown in Figure 4.1, Nexplore divides its interface into four areas: area 1 shows suggested or related queries; area 2 displays thumbnails of retrieved documents; retrieved documents in area 3 are represented by their title, snippet and URL; and area 4 displays sponsored links. Nexplore colours the query and highlights the background of the abstract when the mouse is moved over it.

4.1.3 Topic selection

For each interface, users were given a navigational search task, for which they were asked to find a specific, single correct answer page for a given topic. The topics were chosen to cover areas that were likely to be of interest to young searchers, and where searchers were unlikely to be hindered due to lack of general knowledge about the domain. The three topics were:

A: Find the ARIA chart of the top 50 music singles in Australia (query terms: `top Australia aria`)

G: Find the MSN games website (query term: `msn`)

H: Find the official homepage of the 2009 movie Harry Potter (query terms: `magical potter`)

These topics, and their corresponding answer documents, represent different aspects of navigational searches: the answer for the first topic is a single web page presenting the required (named) information; the second is the hub page for a prime sub-part of the overall MSN website; and the third is the home page (or index) of an overall website.

Figure 4.4: Heatmap showing the gaze of a participant using the Nexplore interface.

4.1.4 Procedure

During the experiment, participants engaged in the following procedure. After reading the explanatory statement, participants were taken through a short calibration session with the eye tracking device. They were then shown a screen that displayed the first search topic. After reading the topic, they clicked a start button to begin the search session. A search result screen with a particular interface was loaded, and the participants were free to interact with the interface however they liked. Once they thought they had found the right answer, participants pressed F10 on the keyboard to move on to the next topic. To keep the participants' eyes focused on the screen, the topic question and instructions were displayed at the bottom of the screen (see Figure 4.4).

One of the main issues with the usability studies is that the experimental condition order can bias the results of the study. For example, users might find using an interface with particular set of topic much easier and faster than other interface. One solution for comparative design is to use a Latin Square to rotate and counterbalancing the condition

Trial	1st task	2nd task	3rd task
1	M- H (4)	C- G (4)	N- A (3)
2	M- G (4)	C- A (4)	N- H (4)
3	M- A (4)	C- H (4)	N- G (3)
4	C- G (3)	N- A (3)	M- H (3)
5	C- A (5)	N- H (5)	M- G (5)
6	C- H (4)	N- G (4)	M- A (3)
7	N- A (3)	M- H (3)	C- G (3)
8	N- H (3)	M- G (3)	C- A (3)
9	N- G (3)	M- A (3)	C- H (3)

Table 4.1: Experimental design.

order effect. A basic Latin square design starts by an $n \times n$ array where n is an treatment that appears exactly once in each row and each column [Kelly, 2009]. Then, a rotation of rows and followed by a rotation of columns take place to ensure there is no encountering between two treatments exist.

Thus, we used a Latin square experiment design with a block of nine trials varying the order in which topics and interfaces were presented to users. Each user was presented with one topic for each interface. Due to some interruptions and other problems, not all combinations were completed exactly the same number of times. Table 4.1 shows the number of times (in parentheses) each of the different combinations of interface (C, M, N) and topics (A, G, H) were completed as the first, second or third task undertaken by one of the users.

4.2 Search interface features

Our experiment involved users interacting with three different search result interfaces that contained different amounts of surrogate text and visual browse features about answer documents on the result pages. In this chapter, we consider the following areas within each

interface page displaying the ranked list of answers.

Surrogate text: Search engines provide surrogates for answer pages from the ranked list of answers. This surrogate text may include the URL of the answer, as well as text from the answer web page title, and a synopsis of the answer web page. The surrogate text for the answer documents is marked (1) on the Middlespot (17%), (2) on the Carrot2 (accounting for approximately 66% of the screen), and (3) on the Nexplore (56%) interfaces.

Browse features: The visual browse features for the answer documents are marked (1) on the Carrot2, (2) on the Middlespot and Nexplore interfaces. Figure 4.2 shows the clustering area in Carrot2, which occupies approximately 19% of the screen. Figure 4.3 shows large images of the answer pages that are displayed in Middlespot and occupying approximately 75% of the screen. Figure 4.1 shows the small region, approximately 7% of the screen, containing the thumbnails displayed by Nexplore.

Other regions: Each interface also had some other regions, including the surrounding screen, banners, and so forth. This accounted for approximately 16% of the screen with the Carrot2 interface, 8% with Middlespot interface, and 37% with Nexplore (since this last interface included a separate area for Wiki Search).

As summarised in Table 4.2, significant portions of the Carrot2 and Nexplore interfaces are given to surrogate text. The great majority of the Middlespot interface, on the other hand, is occupied by visual browse features.

Interface Features	C	M	N
Text features	66%	17%	56%
Browse features	19%	75%	7%
Other regions	16%	8%	37%

Table 4.2: The distribution of interface features.

4.3 Results

We analyse the behaviour of users carrying out the three search tasks using the Carrot2, Middlespot and Nexplore interfaces, based on the relative attention paid to different interface features and the task completion time.

4.3.1 Interface features

Different search interface features attract highly variable amounts of user attention. Figure 4.5 shows the proportions of total viewing time that users spent looking at text, browse and other features for each trial (that is, across all search interfaces and all users). The solid line shows the median time, while the boxes show the 25th to 75th percentiles. Whiskers show the range of the data, with outliers (observations more extreme than 1.5 times the interquartile range) included. Since the time data is not normally distributed ($Shapiro - Wilk, p < 0.0001$), we analyse multi-level factors using the Kruskal-Wallis test, a non-parametric alternative to ANOVA. Pairswise comparisons are made using the Wilcoxon signed-rank test. The relative times for the different features vary significantly ($Kruskal - Wallis, p < 0.0001$). In particular, users spend significantly more time viewing text features compared to browse features ($Wilcoxon, p < 0.0001$) and others ($p < 0.0001$). The difference in viewing patterns between browse and other is not significant ($p = 0.6504$).

Figure 4.6 shows the median time (over all search answer interfaces) users spent looking at different regions of the screen, broken down by cases where users identified the correct or incorrect answer document for each search trial. The text region was the area of the screen that users spent most of their time looking at. Users found slightly more correct answers if they spent a bit more time in this area; on the other hand, when users spent more time looking at the visual browse regions, these proved to be ineffective and could often lead users to incorrect answers rather than correct ones. Time spent looking at both text and browse regions is significantly different between correct and incorrect answers ($Wilcoxon, p = 0.0060$ for text regions and $p = 0.0303$ for browse regions), while the difference is not significant for other areas of the screen ($p = 0.7669$).

Figure 4.7 shows the distribution of the proportion of time that users spent viewing different features, split by the three interfaces. For the Carrot2 and Nexplore interfaces, users spent substantially more time viewing the text features. However, for the Middlespot interface, the browse features (in this case, the screenshots of web pages) attracted the greatest proportion of viewing time.

4.3.2 Task completion time

User task completion performance is evaluated by measuring the time taken to carry out a search task to the users satisfaction. That is: we measure the time from when the search results screen is displayed to the user until the moment that they indicate that they have found a desired answer (generally, by clicking on the hyperlink in the search results list that they choose as their final answer).

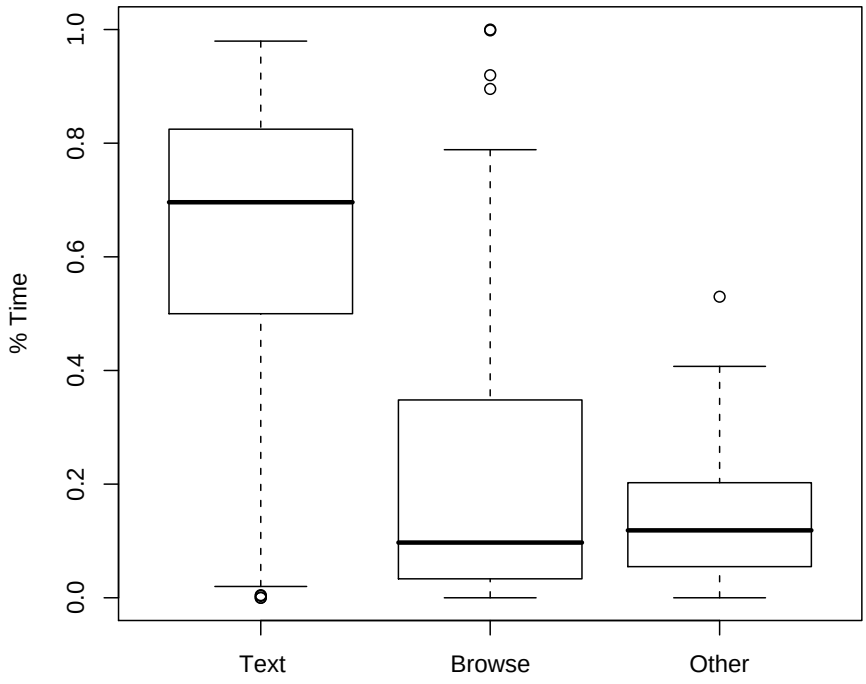

Figure 4.5: Relative time spent viewing different interface regions.

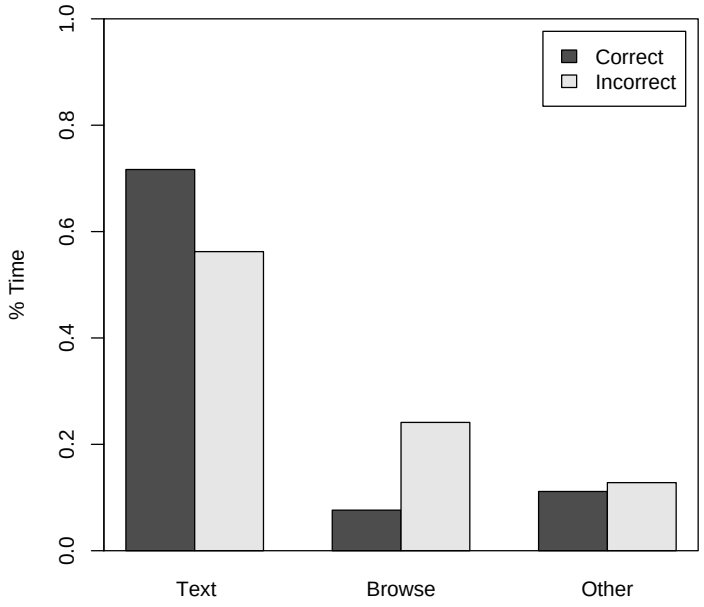

Figure 4.6: Median proportion of time spent viewing different regions for instances when users found a correct or incorrect answer.

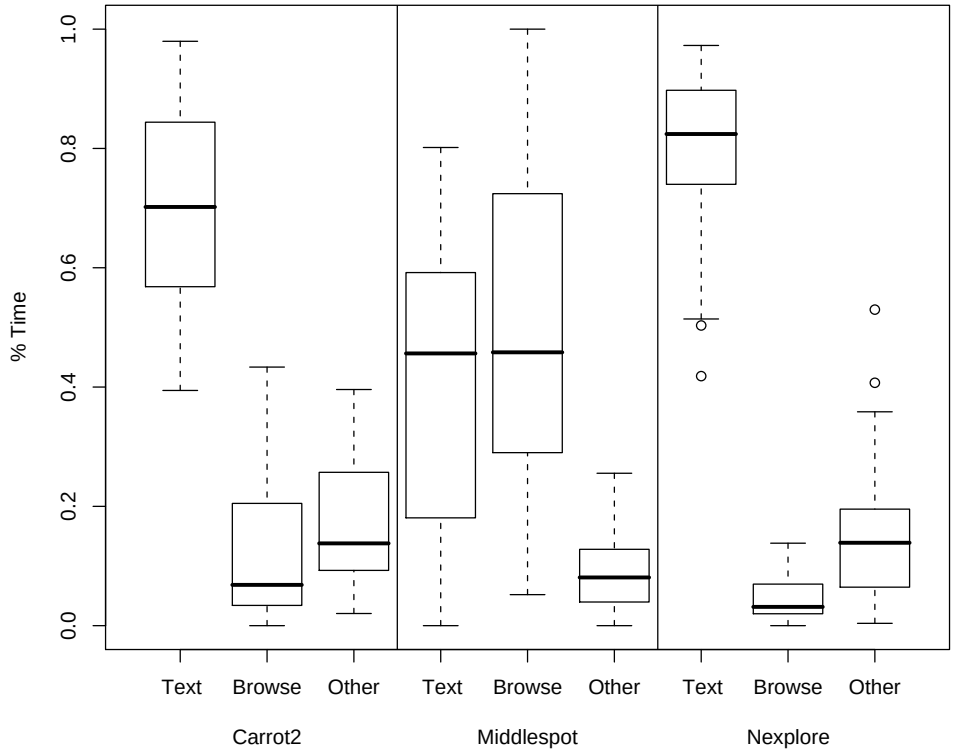

Figure 4.7: Proportion of time spent viewing different components, by interface.

The time data is not normally distributed ($Shapiro - Wilk, p < 0.001$), so we use the Kruskal-Wallis test (a nonparametric alternative to ANOVA) to analyse the significance of multi-level factors, and Wilcoxon signed rank tests for pairwise follow-up analysis.

Figure 4.9 shows the time taken to find an answer, in seconds, for each of the three interfaces. Two outlier points occurred with the Carrot2 interface; these were users who spent additional time browsing the result page, behaviour that was only elicited when using the Carrot2 interface. The use of different interfaces leads to different median search times: 27.53 seconds for Middlespot, 20.02 seconds for Carrot2, and 16.89 seconds for Nexplore. The different interfaces have a significant impact on time ($Kruskal - Wallis, p = 0.048$). In particular, search tasks were completed significantly more quickly using the Nexplore interface, compared to using the Middlespot interface ($Wilcoxon, p = 0.012$). The difference between the other pairs of interfaces is not statistically significant: Carrot2 and Middlespot ($Wilcoxon, p = 0.176$); and Nexplore and Carrot ($Wilcoxon, p = 0.310$).

Figure 4.8 shows median time, split by interface and search topic. It can be seen that different search topics cause some variation; in particular, different interfaces appear to offer advantages and disadvantages for different topics. For example, Harry Potter is the topic that requires the longest time to resolve with the Carrot2 and Nexplore interfaces; however, with the Middlespot interface, the ARIA topic is the slowest. Overall, the search topic does not have a statistically significant effect on time ($Kruskal - Wallis, p = 0.127$). Similarly, while some variation in task completion time is observed between users (with some users being relatively slower or faster than others across all three topics), the difference between users is not significant ($p = 0.053$).

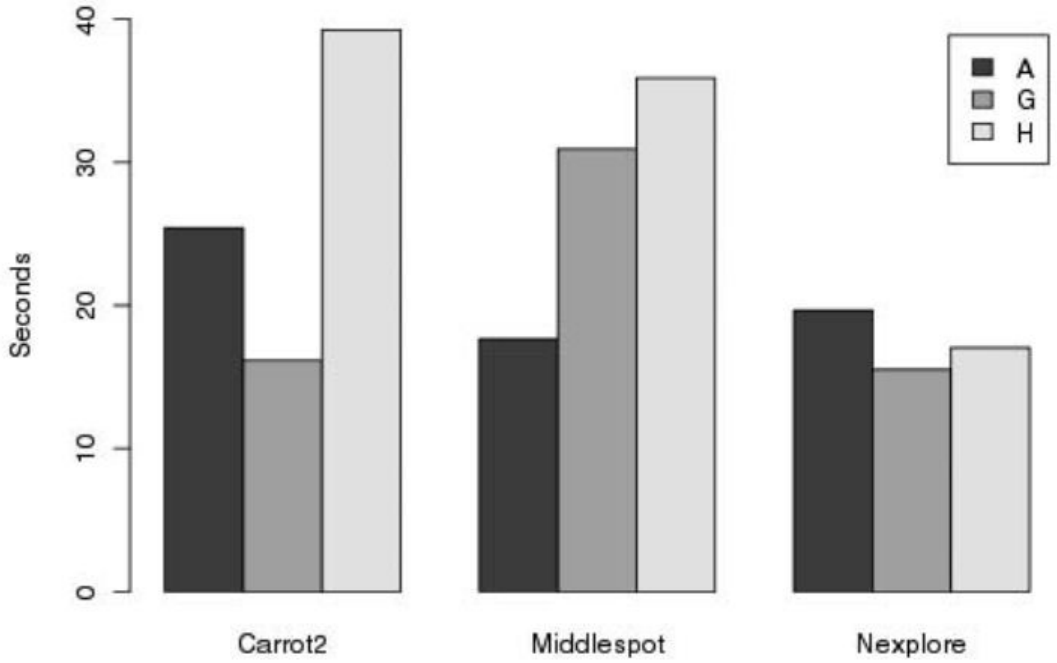

Figure 4.8: Median time spent on the three web search interfaces, divided by topics A, G and H.

We have also investigated the rate of incorrect responses. Table 4.3 shows, for each search session, how many users failed to identify the correct answer resource. The numbers in brackets in the table indicate the total sessions by interface, hence the combined total is 96 sessions. The error rate when using the Middlespot interface is substantially higher than when using either of the others.

Topic	**C (33)**	**M (32)**	**N (31)**
A	33%	90%	22%
G	0%	25%	20%
H	45%	20%	25%
Mean	26.6%	45%	22.3%

Table 4.3: Percentage of search sessions where participants did not find a right answer. The numbers in brackets are the total number of sessions per interface.

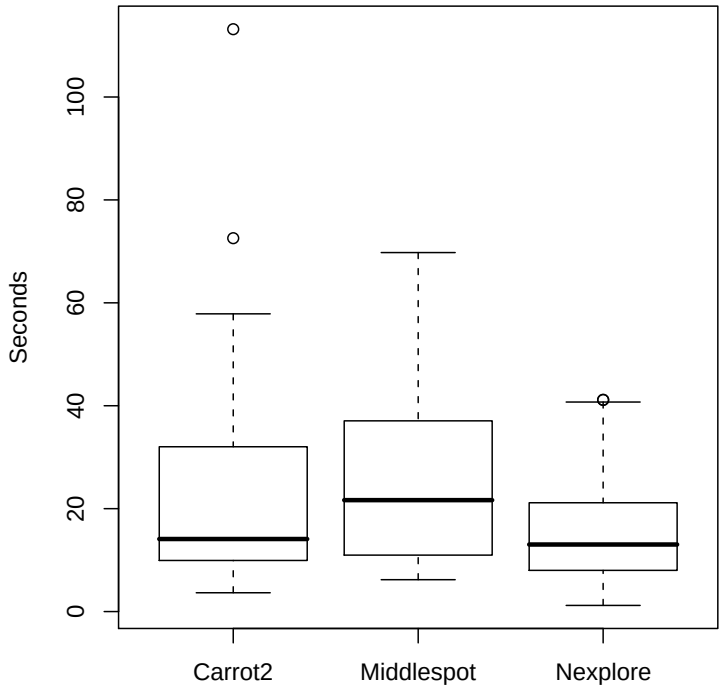

Figure 4.9: Task completion times by interface.

Variation can also be introduced by other sources. The effect of using different search topics was significant ($Kruskal-Wallis, p = 0.0330$). Moreover, because we used real search interfaces and live search results, the rank of the correct answer items in the search results lists from different interfaces varied somewhat. Although the ranks were similar on average (rank 7.6 for Carrot2, 6.3 for Middlespot, and 6.0 for Nexplore) this did have a significant effect on task completion time ($Kruskal - Wallis, p = 0.0048$). The different users participating in the experiment were not a significant source of variation ($Kruskal - Wallis, p = 0.1227$).

However, this analysis includes all user responses, irrespective of whether the user actually found the correct answer required for the query. (We investigate this issue next.)

Answer	Carrot2	Middlespot	Nexplore
Correct	24	18	24
Incorrect	9	14	7

Table 4.4: Distribution of correct answers by interface.

4.3.3 Search success

Users were asked to indicate when they felt that they had found the correct answer to the query. However, in many cases users did not in fact identify the correct resource. Table 4.4 shows the number of incorrect and correct answers found, split by the interface used. The results are strongly indicative of higher success rates with both the Carrot2 and Nexplore interfaces (72.7% and 77.4% of answers are correct, compared to 56.2% for Middlespot). However, the differences are not statistically significant ($Fisher, p = 0.1746$).

We re-analyse the time taken for task completion, using only those trials for which users identified the correct resource in response to the information needed. For these responses, the difference between interfaces is greater, and statistically significant ($Kruskal - Wallis, p = 0.0077$). Differences between the interfaces on a pairwise basis are also more pronounced: the median task completion time with Middlespot at 23.71 seconds is significantly longer than that for Carrot2 at 12.81 seconds ($Wilcoxon, p = 0.0112$) and for Nexplore at 12.21 seconds ($Wilcoxon, p = 0.0027$). The difference between Carrot2 and Nexplore is not significant ($Wilcoxon, p = 0.7360$).

Moreover, when considering only those results where users successfully identified correct answers, the effects from topic and user variation are not significant (Kruskal-Wallis, p = 0.3445 and 0.2743, respectively). The rank of the answer item only has a weakly significant

effect ($Kruskal - Wallis, p = 0.0619$).

4.4 Discussion and summary

Search result interfaces are an important component of information retrieval systems, and can have considerable impact on overall search task performance. In this chapter, we have analysed three publicly available search interfaces, and examined how user attention is split between various features that the search providers make available.

One of the main features of Nexplore is that results are presented using a mixture of text and visual aids. The use of colours and screenshots for the results can help users to identify relevant information. Figure 4.4 shows a screenshot from the eye tracking analysis for the Nexplore interface, using the ARIA chart search topic. Regions of interest (those that received more frequent and longer gazes) are highlighted. It can be seen that the participant focused on the textual summary (snippet, title, and URL), and briefly looked at the thumbnail image of the correct answer before clicking on it. The participant also referred back to the search topic shown at the bottom of the screen. In comparison, the Middlespot interface devotes relatively less space to textual information, with most screen space being used to display screenshots. Moreover, the dynamic zooming features that are activated by mouse-hovering make the Middlespot interface more difficult to use. The Carrot2 interface, on the other hand, presents results in a completely text-based manner. Although the interface enables the clustering of results, the use of this feature was rare. Based on the timing of results, task completion time for the Carrot2 appears to fall between Middlespot and Nexplore.

Additionally, our analysis has shown that users spend significantly different proportions

of time interacting with text, browse and other components of the interfaces. Not surprisingly, these proportions differ between the three interfaces; for Nexplore and Carrot2, text is preferred, while for Middlespot (which presents much less text to the user), browsing features are viewed for a longer proportion.

We have also analysed how task completion time differs between the interfaces, and examined success rates in identifying correct answers for given information needs. The results show that users spent a significantly longer time interacting with the Middlespot interface and found the fewest correct answers. We conclude that, for the navigational search tasks, text features are important in guiding users to finding correct answers quickly.

For the small sample of named-resource finding search tasks, it appears that text information can be vital in supporting users to find the answers that they need. Whether this would also apply to other search tasks, such as informational tasks, will be the subject of future research.

Chapter 5

Evaluating visual summaries for informational search

Recent studies have developed various novel approaches to visual summaries, aiming to improve the effectiveness of search results. In this chapter, we evaluate the effectiveness of additional visual summaries (visual snippets, visual tags, excerpt images, and thumbnails) on web search interface using informational topics. We also investigate the impact of these visual summaries on user seeking behaviour and performance. Each one of these visual summaries is presented on an interface together with a text summary.

This chapter investigates our second research question for information search: Does providing additional visual summaries for the presentation of web search results impact on users' information-searching behaviour and performance? We investigate particularly the following sub-questions:

1. To what extent does providing additional visual summaries for the presentation of web search results help users to predict relevant answers for informational topics?

2. To what extent does presenting additional visual summaries impact on user information seeking behaviour, particularly on how users interact with the text summaries for informational topics?

3. To what extent does the presence of additional visual summaries affect the cognitive load of users, particularly the effort needed to answer the given informational search topics?

Our analysis shows that users spend significantly less time looking at textual summaries when visual summaries were available. However, overall, the results suggest that visual summaries do little to increase user performance with informational topics.

The chapter is organised as follows: we describe our experimental design including the visual summaries, users and topics in Section 5.1. Experimental results are analysed in Section 5.2, and discussion and summary are presented in Section 5.3.

5.1 Experimental framework

In order to evaluate the effectiveness of different approaches for visual summaries, and study the impact of these visual summaries on user seeking behaviour and performance, we conducted a user study that involved a series of five informational search topics using different search interfaces where visual summaries are a primary component of the search results presentation.

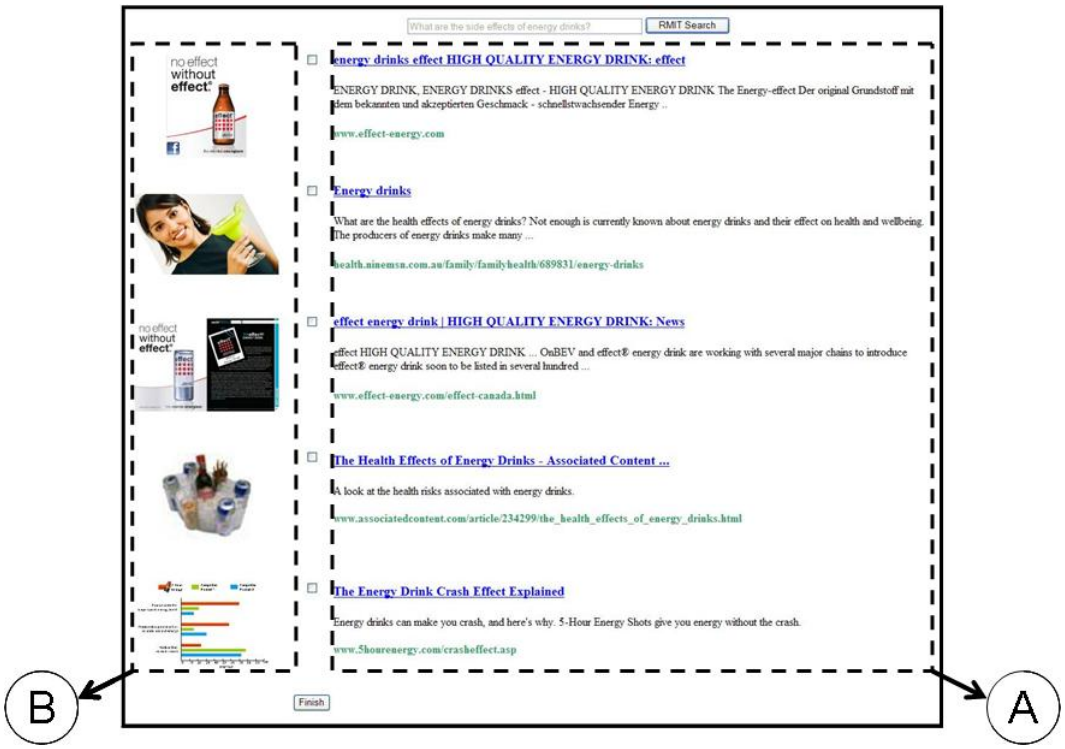

Figure 5.1: Salient image interface: (A) Text summary region. (B) Visual summary region.

5.1.1 Experimental setup

In our user study, the participants were mostly undergraduate and high school students with some interest in computer science visiting RMIT University at the 2010 Open Day. A plain language statement was given to the subjects to outline the purpose of the experiment, the procedure, the tasks to be performed, and the data to be collected. Based on this information, 65 participants chose to take part in the experiment. However, due to interruptions and difficulty with calibrating the eye tracking for some volunteers (we eliminated users with less than 80% capture accuracy), the collected data of only 50 participants is included in the analysis. A short oral presentation about the visual summaries was given to each participant, but no training was given on the interfaces to be used.

5.1.2 Interfaces

Five interfaces were evaluated: one text-only interface and four visual interfaces, where each visual interface presented a different approach to visual summaries: thumbnail, visual tag, salient image and visual snippet. The four visual summaries were chosen based on different parameters. Thumbnail was chosen in terms of popularity, while visual snippet and salient image were chosen based on their effectiveness as shown in certain studies [Li et al., 2008b; Teevan et al., 2009; Loumakis et al., 2011]. Visual tag is another approach to investigation. A textual summary was presented in each interface and consisted of: a web page title, a text snippet (that is, a brief textual extract designed to relate the query terms to quotes from the source web page), and the URL of the underlying web page. Apart from the text-only interface, each of the other interfaces included an additional visual summary for each search result item.

To minimise presentation variations among the five interfaces and to focus our study on visual and text summaries, we designed a template in order to present summaries consistently and uniformly across the five interfaces. We also tried to keep our testing interfaces similar to that of main stream web search engines, as several studies have shown that users are more comfortable with a familiar presentation of search results [Hoeber and Yang, 2006; Micarelli et al., 2007; Joho and Jose, 2008]. Previous studies have indicated that presenting visual summaries on the left of text summaries is more recognisable for English language users [Mishkin and Forgays, 1952; Bryden and Rainey, 1963; Fontenot, 1973].

Therefore, in the template, we presented text summaries on the right and the visual summaries on the left. The visual summaries were displayed with a maximum of 200×150

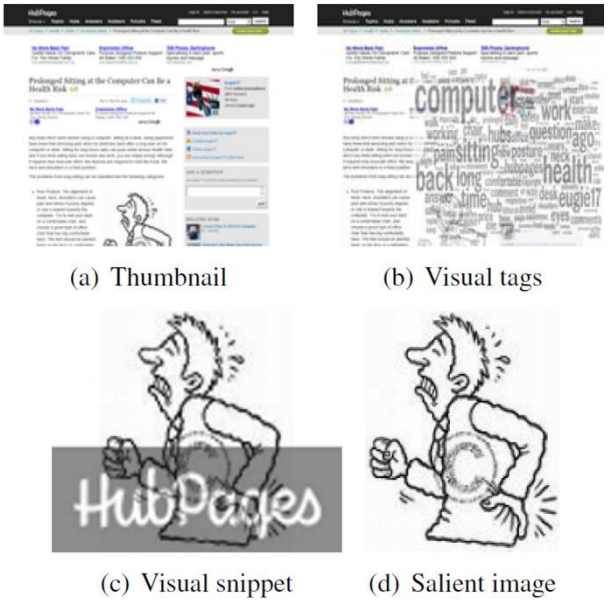

Figure 5.2: Examples of the four types of visual summaries.

pixels, while keeping the original image ratio [Kaasten et al., 2002; Won et al., 2009]. For each topic, the five interfaces presented exactly the same textual surrogates (title, snippets and URLs) in exactly the same place with the format discussed above. Visual summaries were also displayed in exactly the same place for all the topics and for all the interfaces, except for the text-only interface where a white space replaced the visual summaries.

5.1.3 Types of visual summaries

Thumbnail (Thum): A thumbnail is the screen shot of a web page, as shown in Figure 5.2 (a). This is the most common visual summary that has been studied in previous work, and is presented in commercial search engines, such as Bing and Google, if a user hovers the mouse over a textual answer summary. In this chapter, we will refer to this thumbnail as a plain thumbnail. In order to control the properties of thumbnails in the experiments, a software

tool called WebShot[1] was used to create the snapshot for the test collection.

Visual Tag (Tag): A tag cloud shows the most frequent words of the source document using different font size and colour to indicate the importance and repetition of terms. A tag cloud is a popular method for showing web search results [Bateman et al., 2008; Sinclair and Cardew-Hall, 2008; Schrammel et al., 2009]. A visual tag cloud is an approach which combines the snapshot of the retrieved web page with its tag cloud. The key idea is that the combination of the snapshot and tag cloud can provide the user with effective cues about the content of the document source. The construction of the visual tags includes two main stages. Firstly, a transparent image of each tag cloud was created using Wordle website2. The next step was to combine this with the thumbnail of the related web page. Buscher et al. [2009] have found that people focus more on the top left corner of a web page, because the logo and the main navigation bar are usually located in that area, so to preserve this information region the tag cloud was located on the right of the thumbnail as shown in Figure 5.2(b).

Salient image (Img): Li et al. [2008b] crawled three websites (MSN.com, MIT.edu and CNN.com) to evaluate an interface using image excerpts, dominant images that are relevant to the user's query along with text summaries, compared with a text-only interface. The results showed that the image excerpt together with the text summary helped the user to find answers in less time than with the text-only interface. Example for the salient image is show in Figure 5.2 (d). The same Google image search was used to extract a salient image, as explained for the visual snippets.

Visual snippets (VSnip): Teevan et al. [2009] developed a visual snippet consisting of

[1] http://www.websitescreenshots.com

three components: a page title, a salient image and a logo of a website. The salient image is the most relevant image from the source document for the given query. If the salient image is not available, a snapshot of the web page is taken instead, while the logo is ignored, if not identified. In our study the visual snippets (VSnip) consisted of the website logo and a salient image from the retrieved web page as shown in Figure 5.2 (c). The page title is already displayed in the related textual surrogate, so it was not repeated in the visual snippet. To obtain the image to use, we ran a Google image search over the target URL and selected the top-ranked image.

5.1.4 Topic selection

In this chapter, we focus on informational search tasks that aim to find specific information for a given topic. Five informational topics on general knowledge were developed :

1. What are the side effects of energy drinks?
2. What is a gecko?
3. What is an appropriate sitting posture at a computer?
4. What is a solar eclipse?
5. What is a Vuvuzela?

To obtain realistic search engine results, the top ten search results from the Bing search engine were selected for each query. Wikipedia entries were excluded as they have obvious answers for the experimental tasks, then five items were randomly chosen from the remaining search results. To ensure balance in the result sets, the quality of the five selected items was

restricted to include at least one relevant and one non-relevant answer, as judged by the authors.

5.1.5 Procedure

After reading a topic from the screen, the participant clicked a start button to load the search interface. Five items were displayed as search results, and participants were asked to select all items that they consider to be a relevant answer for the task. Then, the participant clicked on a finish button to move to the next task.

The presentation of topics and interfaces were determined by a Latin square, described in Section 4.1.4, giving 25 combinations of interfaces and topics, to control for presentation order effects.

5.2 Results

We analyse user behavior when carrying out the five informational search topics, using a different interface for each, based on topic completion time and the relative attention paid to different summary features (page title, textual snippet, URL and visual summary).

5.2.1 Effectiveness of relevance prediction

In the user study, participants were asked to select all answer items that looked relevant for the given search topic. Table 5.1 shows the Click Precision, Click Recall and Click F-measure for how effectively the users were able to identify relevant answers.

Although average Click Precision indicates that the tag cloud can mislead users, F-test

Measures		Text	Thum	Tag	VSnip	Img
Click Precision	Average	0.865	0.794	0.689	0.800	0.820
	Stdev	0.237	0.270	0.365	0.322	0.298
Click Recall	Average	0.582	0.600	0.505	0.592	0.535
	Stdev	0.263	0.316	0.320	0.335	0.303
Click F-measure	Average	0.645	0.625	0.545	0.624	0.569
	Stdev	0.190	0.232	0.290	0.271	0.250

Table 5.1: Click Precision, Click Recall and click F-measure for user selection of search result items.

Answer	Txt	Thum	Tag	Img	VSnip
Relevant	71	71	64	64	72
Non-relevant	18	26	34	21	27

Table 5.2: Distribution of the number of relevant and non-relevant answers selected by users, grouped by interface.

shows no significant difference between the interfaces ($F = 2.3009, p = 0.0593$). Thumbnail interface achieves the highest average Click Recall (0.6), an F-test shows no significant differences between the interfaces on Click Recall ($F = 0.8796, p = 0.4767$). Additionally, the results showed no significant differences between the interfaces on click F-measure ($F = 1.2337, p = 0.2971$).

The number of relevant and non-relevant items that users selected are shown in Table 5.2, split by the interface used. The results show largely consistent rates of selection of relevant answers with no significant differences between the interfaces ($\chi^2, p = 0.9168$). The results also show that users selected fewer non-relevant answers using the text-only interface, compared with the visual interfaces, while users selected fewer non-relevant answers with the image interface in comparison with the other visual interfaces. However, a Chi-squared test shows no significant difference between the interfaces on the total number of non-relevant

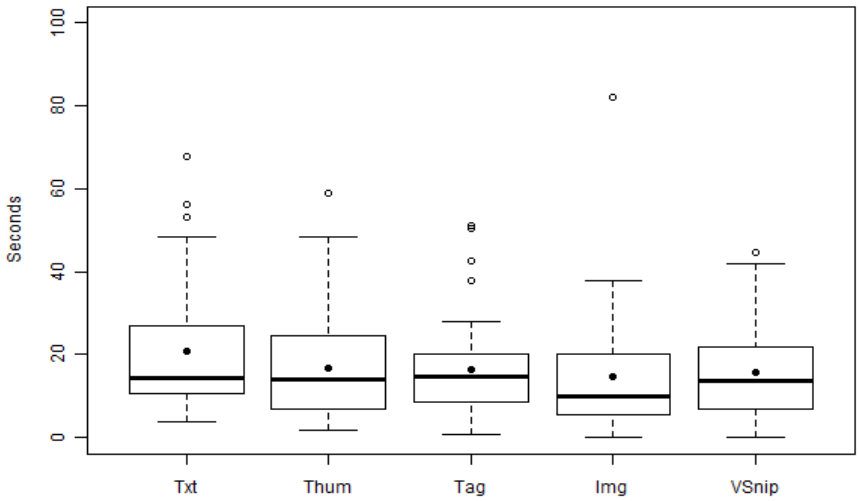

Figure 5.3: The time in seconds spent viewing text summary regions.

answers ($p = 0.2003$).

5.2.2 Interaction with textual summaries

User interaction with the search results was captured using eye tracking data. By using this dataset, we can investigate how users interact with textual summaries when additional visual summaries are presented. Figure 5.3 shows the amount of time spent looking at the text summary regions. Users spent substantially more time looking at the text region for all interfaces. While users in general spent less time looking at the text region when using a visual summary interface, particularly with image interface, F-test show no significant difference between the interfaces ($F = 1.5694, p = 0.1831$). That is, presenting the visual summary feature decreased the attention that users gave to the text-based summary information.

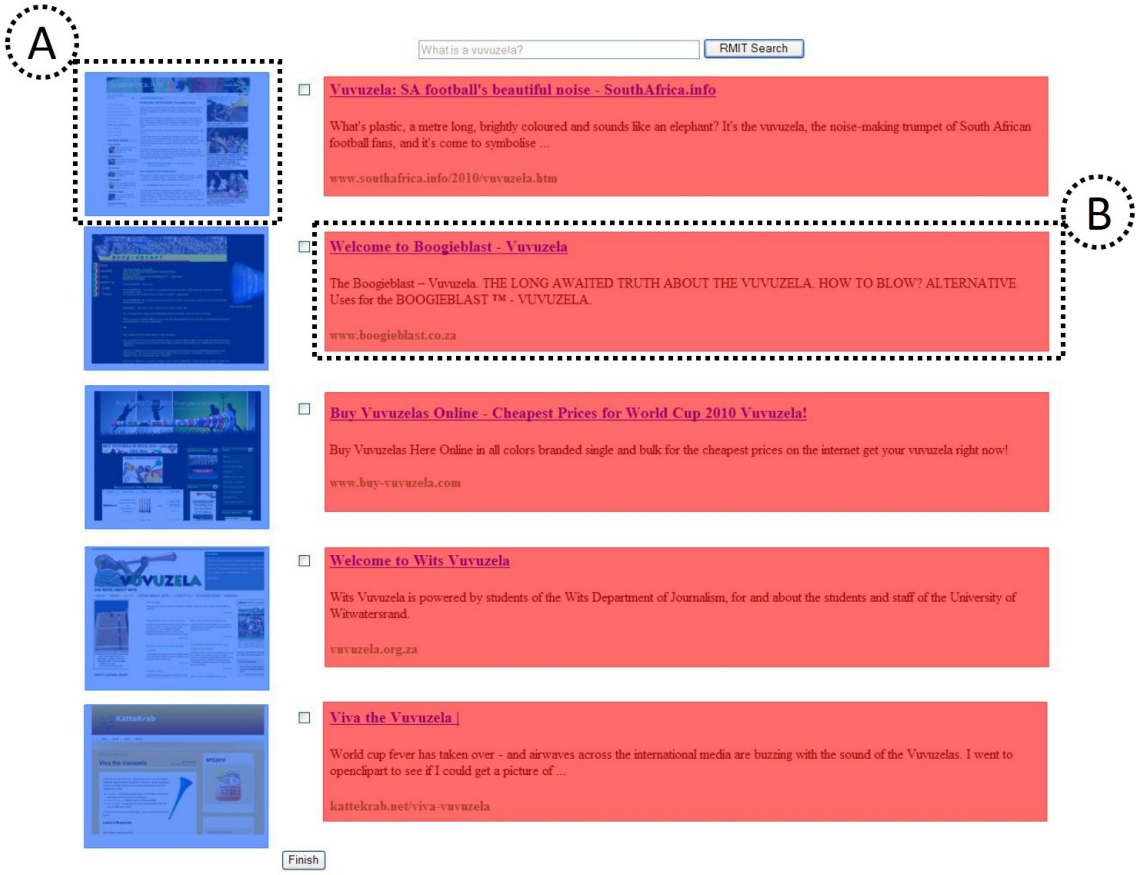

Figure 5.4: The mask used to collect time spent on informative components: (A) visual summary. (B) Textual summary.

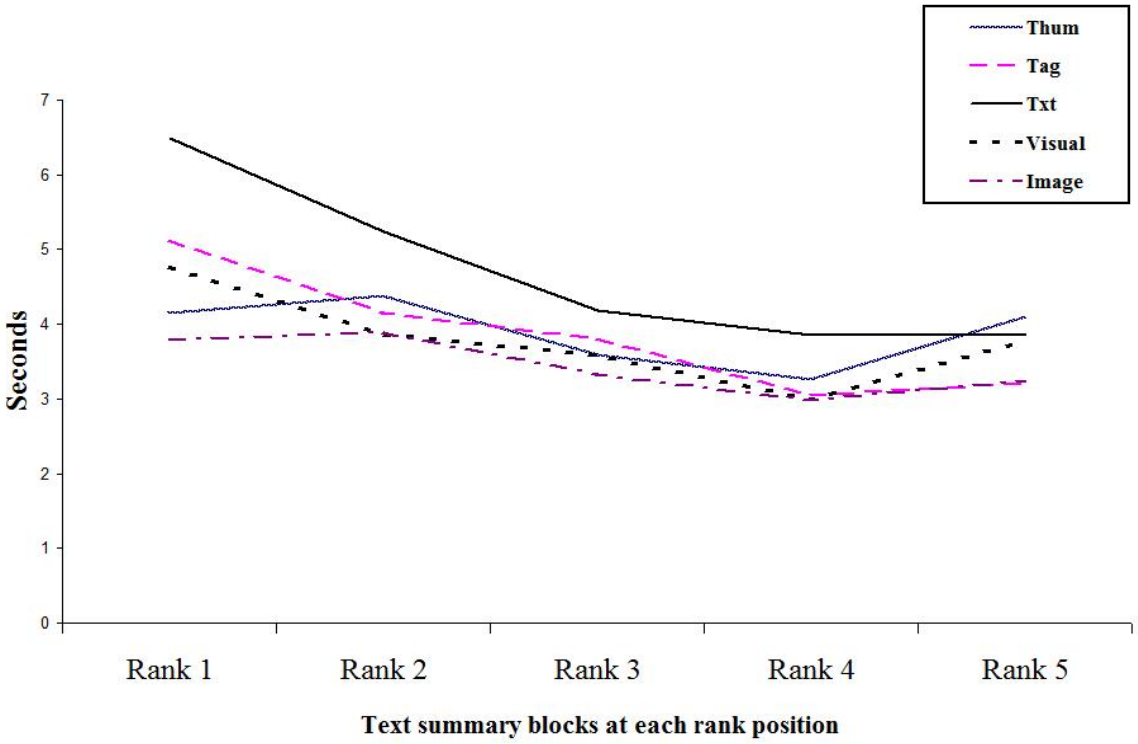

Figure 5.5: Average time spent on the textual surrogates for the five search result items.

We also analysed how users scanned the search results with each interface by collecting the time that users spent looking at each text search result item. The gaze regions were bounded on the interface component, leaving regions of white-space between them, as shown in Figure 5.4. The results show that users were more influenced by the vertical list of the search results when they used the text interface, but this behaviour was less apparent on the interfaces that present visual summaries as shown in Figure 5.5. Users with text-only interface paid more attentions to the top ranking items and less attention viewing the items on the bottom of the search results list. In the visual interfaces, user attention distribution was more balanced across the items of the search results list.

5.2.3 Interaction with the visual search interfaces

Next we study user attention in relation to the different information summaries (visual and visual regions). A broader comparison was conducted, by pooling the data for the four interfaces that include visual summaries, and then comparing the two attention areas: visual and text regions. The results show that users spent significantly more time looking at the textual summaries than on the visual summaries ($p < 0.0001$). Additionally, we evaluate the time spent on visual summaries between the four interfaces by comparing the time spent on visual summaries on each visual interface. However, an F-test show no significant different between the four interfaces ($F = 1.709, p = 0.1665$).

5.2.4 Overall task completion time

The performance of users to complete a task was evaluated by collecting: the time users spent on each task; the time taken to first selection; and, the time taken to select first relevant item. However, no statistically significant differences were found between the five interfaces. Figure 5.6 shows the total time spent to answer the search tasks for each interface. Although users required the least amount of time to finish their search tasks with the salient image interface, F-test shows no significant differences ($F = 0.3996, p = 0.8088$). Also, we calculated the average time that a user spent to answer each search task for each interface. No statistically significant difference was observed between the interfaces on the measures of time completion.

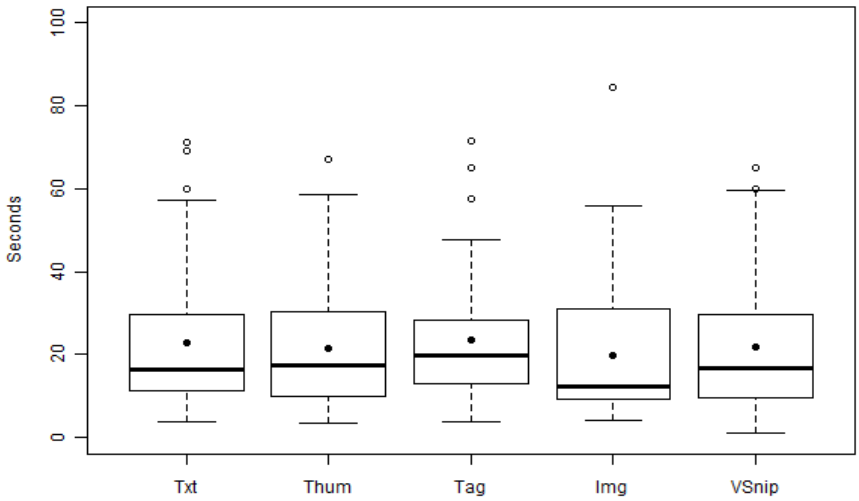

Figure 5.6: The total time spent to finish search tasks for each interface.

Interface	Txt	Thum	Tag	Img	VSnip
Uniquely viewed summaries (text or visual summaries)	246	245	245	247	241
Uniquely viewed text summaries	246	232	228	224	234

Table 5.3: The total number of uniquely viewed items split by interface.

5.2.5 Viewing and reviewing search result items

Analysing user viewing behaviour of text summaries can provide indications about user mental effort. Figure 5.4 shows the mask used to collect, for each search topic, the total number of uniquely viewed items and the total number of times that users viewed search result items. For the visual interfaces, an item was counted as viewed if the user viewed one or both (textual and visual) summaries of that item.

We collected the uniquely viewed items (out of a possible 250 search result items to

Interface	Img	Tag	Thum	VSnip
Tag	**0.0140**	-	-	-
Thum	0.1691	0.8781	-	-
VSnip	0.4160	0.5939	0.9867	-
Txt	**0.0003**	0.8458	0.2851	0.1001

Table 5.4: The results of Tukey's HSD test for pairwise comparison of the percentage of re-viewing for the entire time required on answering the search topics.

finish answering the given search topics for each subject) for each interface, as shown in Table 5.3. No significant difference was found between interfaces ($\chi^2, p = 0.9991$). In addition, we collected the total number of uniquely viewed text summaries (total number of single viewed items) by each user, split by interface, as shown in Table 5.3. The difference between interfaces is again not significant ($\chi^2, p = 0.8799$).

To analyse user mental effort with search results in more detail, we calculate the percentage of re-viewing (as described in Section 3.4.5) to finish a search task in each session. The percentage of re-viewing measures the difference between the total and unique views of search results.

An F-test shows a statistically significant difference between the interfaces ($F = 5.0147, p = 0.0007$). A Tukey's HSD test was then used for multiple comparisons between the five interfaces, and results are shown in Table 5.4. Users spent significantly less effort (re-viewed fewer search items) when using the salient image interface compared with the text-only interface ($p = 0.0003$) and the tag interface ($p = 0.0140$). The results also show that, apart from the salient image interface, no significant difference was found between the three visual interfaces on the required effort when comparing with text-only interface. Thus, salient image summaries help a user to spend less effort to find relevant answers for informational topics.

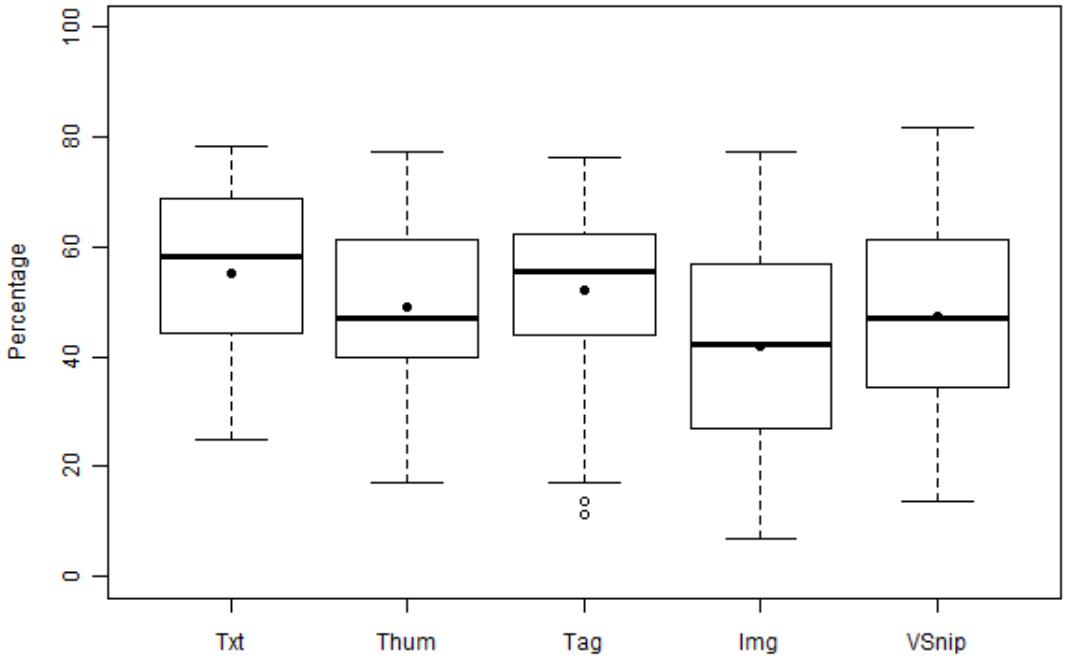

Figure 5.7: The percentage of re-viewing search result items for the entire time required to answer search topics split by interface.

5.3 Discussion and summary

In this study, we evaluated the impact of different types of visual summaries on user behavior and performance. Fifty participants carried out five informational topics using five different interfaces. Our study primarily focused on evaluating the ability of users to predict the relevance of answers when visual summaries are provided. Other studies [Jiao et al., 2010a; Teevan et al., 2009] focus on evaluating visual summaries in terms of finding and re-finding issues, whereas in our study we considered informational search tasks.

Providing additional visual summary information with the text search results did not significantly improve the ability of users to predict the relevance of a result page for an informational search task. Although the thumbnail interface achieved the highest average Click Recall, a statistical test showed no significant difference between the interfaces. Also, no significant difference was found for the number of relevant result pages that users selected for each interface. Further, the results show that adding visual summaries may mislead users to select non-relevant results pages for the search topics. A possible reason for explaining this behavior is that users are not as familiar with these novel approaches as with standard text-only result lists.

We studied user behavior when presented with an additional visual summary, and the results show that visual summaries significantly affect user behavior. Although the informational search tasks seem to require reading text more than looking at pictures, users in general spent less time looking at the text region when using a visual summary interface. This may also explain the lower performance when predicting the relevance of answers items when visual information is displayed.

Users spend more time looking at tag clouds, but there is no significant difference in attention between the four interfaces that include visual information. Also, the results show that users scan the search results exhaustively for the text interface, but economically for the visual interfaces. With the text interface, users spent more time looking at the top items and this amount gradually decreased as they move down the ranked list, while for the visual interfaces, the amount of time per item shows less variation.

Furthermore, we collected the number of uniquely viewed items and number of times users viewed text summaries, but no significant difference was found between the interfaces. However, the results outlined significant differences between text-only interface and the image interface on the re-viewing percentages of search result items. This suggest that salient image help users to spend less mental effort to find relevant answers for informational topics.

In addition, we collected the time users spent on each task, time taken to first selection and time taken to select first relevant item. However, given our sample size no statistically significant differences were found. This suggests that, apart from salient image summaries, visual summaries do not provide enough information for informational search tasks, since the answers for this type of search are more likely to be located in the text rather than visual summaries.

Chapter 6

Evaluating visual summaries for navigational search

In the previous chapter, we evaluated the effectiveness of four visual interfaces (thumbnails, salient images, visual snippets and visual tags) and a text-only interface using informational topics. In this chapter, we evaluate the impact of the four types of visual summaries with navigational web searches − where the user is looking for a specific resource − on user information seeking behaviour and performance. Five interfaces were designed: a text-only interface and four visual interfaces. Each visual interface presents a different approach from adding visual summaries (thumbnails, visual tags, excerpt images and visual snippets) along with the text summaries.

This chapter aims to address our second research question for navigation search: Does providing additional visual summaries for the presentation of web search results impact on users' information-searching behaviour and performance? Particularly, we investigate the

following sub-questions with navigational topics:

1. To what extent does providing additional visual summaries for the presentation of web search results help users to predict relevant answers for navigational topics?

2. To what extent does presenting additional visual summaries impact on user information seeking behaviour, particularly on how users interact with the text summaries for navigational topics?

3. To what extent does the presence of additional visual summaries affect the cognitive load of users, particularly the effort needed to answer the given navigational search topics?

4. To what extent do additional visual summaries impact on users' browsing strategies and hence their search effectiveness with navigational topics?

Our analysis shows that how users interact with visual summaries significantly affects user performance and behaviour on navigational search topics. Some types of visual summaries not only significantly help the user to find answers in a short amount of time compared to a text-only baseline, but also significantly reduce the amount of effort required for the process of extracting and understanding the information from the search result page. Our analysis also shows that different amounts of attention spent on visual summaries corresponds to different forms of browsing on the search results page.

The chapter is organized as follows: our experimental design, including the types of visual summaries, users and topics are explained in Section 6.1. Section 6.2 presents the

analysis of our experimental results. The discussion and summary of the results is presented in Section 6.3.

6.1 Experimental framework

The purpose of this study was to investigate how five representations for summaries of search results impact on the users relevance prediction for specific navigational search tasks. To examine this aim we conducted an experiment that involved fifty participants. Each participant carried out five navigational search topics using each representation. This section introduces the experimental setting.

6.1.1 Experimental setup

Participants were undergraduates and high school students, who were mostly visitors at RMIT University Open Day in August 2011. All had some interest in computer science. A one-page plain language statement was provided to outline the purpose of the experiment, the type of tasks to be performed, the procedure and the nature of the data that would be collected. Based on this information, more than 60 participants were recruited. No formal training was given about how to use the interfaces involved, but a short oral presentation introducing the different visual summaries was given to each participant. Due to interruptions and calibration issues with the eye tracking, we only used the data collected from 50 participants for analysis.

Participants were asked to answer a series of five navigational search topics using different interfaces. The topic search results were represented by five fixed items and participants

selected an answer using a mouse. Browsing the actual web pages embedded in the text search results was disabled, so no interacting search was undertaken by the participants. Participants were instructed to rely solely on the information presented on the given search results page, to predict the relevant answers. Participants were asked to select only one relevant answer.

Experimental data were collected using the Tobii T60 eye tracker. The five candidate summaries were presented on a single non-scrolling page, so participants did not need to have their visual attention distracted by needing to scroll around the search results page. An example of the layout is presented in Figure 5.4, in Chapter 5.

6.1.2 Interfaces

Five interfaces were designed, (text-only, thumbnail, visual tag, image and visual snippet), discussed in Chapter 5. For the same topic, all interfaces had exactly the same text summaries (for example, B is the text summary for the second answer). However, the visual summaries (for example, A is the thumbnail for the first answer) were changed for each of the four different approaches to visual summaries (or no visual summary for the text-only interface) as described in the next section.

6.1.3 Types of visual summaries

Four types of visual summaries were designed as discussed on Chapter 5. A **Thumbnail (Thum):** A thumbnail is the screen shot of a web page, as shown in Figure 6.1 (a). A **Visual Tag (Tag)** is shown in Figure 6.1 (B), and **Salient image (Img)** is show in Figure 6.1 (d).

Visual snippets (VSnip)is the combination of the website logo and a salient image from the retrieved web page as shown in Figure 6.1 (c).

6.1.4 Topic selection

In this chapter, we focus on navigational search tasks, that aim to find a single specific correct answer page. The topics were chosen to cover general areas of interest so as to avoid users needing any specific domain knowledge. Table 6.1 shows the five navigational search topics developed for this study. The tasks specifically identified the unique target resource that the user was supposed to find. The different topics were designed to meet the participants expected areas of interest and knowledge, while also covering different aspects of navigational search, such as finding the homepage or a single particular webpage. For each topic, we sent its corresponding query terms to the Bing search engine (c.f. Table 6.1) and then chose five candidate summaries, from the top ten search results, for presenting to a participant. We made sure that there was only one correct answer from the five presented summaries.

6.1.5 Procedure

After reading a topic, a participant would click on a start button to load the search results with a particular interface. The participant could then interact with the search result screen to select the one answer they believed represented the relevant answer page. Then the participant clicked on a finish button to move to the next task.

We used a Latin square experiment design, described in Section 4.1.4, in which each interface and topic was rotated in each position and resulted in 25 combinations for a complete

Topic	Query terms	Aspect	Domain
Find the Facebook home page.	Facebook	Home page	Social
Find the ARIA web page of the top 100 albums for 2010.	top 100 albums for 2010	Single web page	Music
Find the Optus web page for the Apple iPhone Mobile Phone offer.	Optus iphone	Sub-part of the overall Optus website	Shopping
Find the iTunes trailers for the film "Harry Potter and the Deathly Hallows" Part 2.	"Harry Potter and the Deathly Hallows" Part 2	Single web page	Movies
Find the Microsoft biography web page for Bill Gates.	Bill Gates	Sub-part of the overall Microsoft website	People

Table 6.1: The five navigational search topics involved in this study.

design. To minimise variation among participants, each sequence of the 25 combinations of interfaces and topics was run twice.

6.2 Results

In this section, we describe our experimental investigation into the impact and effectiveness of presenting additional visual summaries for web search with navigational search topics. The analysis is based on topic completion times, the attention paid to the informative components of the interfaces, the participants' responses to given questionnaires, and the participants information search patterns through monitoring participants eye movements.

Interface	Txt	Thum	Tag	Img	VSnip
Correct	33	40	39	34	37
Incorrect	17	10	11	16	13

Table 6.2: The total number of correct and incorrect answers selected by users split by interface.

6.2.1 Effectiveness of relevance prediction

In this study, participants were asked to select a single relevant answer from five candidate summaries, for a given search topic and a given interface. As discussed in Section 6.1.5, there are 250 search sessions in total. Table 6.2 shows the total number of correct and incorrect choices made by users, split by interface. The results show that users selected more correct answers using the visual interfaces, compared with the text-only interface. However, a Chi-squared (χ^2) test shows no significant difference between the interfaces on the total number of correct answers ($p = 0.9073$).

6.2.2 Task completion time

We divided task completion time into three components: the time taken to select an answer; the total time required to finish the task; and, the time required after selecting the answer before moving on to the next task. This last component reflects extra user effort required after finding an initial answer, for example, examining other items to increase confidence in their choice. The loading time of the search results page was excluded from the analysis.

Table 6.3 shows the average time taken to select an answer, split by interface. The time spent on the text-only interface is longer than that on other interfaces, showing that a user generally needed extra time when browsing a text-only result page. Across all users an F-test

Interface	Txt	Thum	Tag	Img	VSnip
Average time required to select an answer	12.0444	7.8711	7.3869	9.0838	7.0268
Average time required from selection to the end of search session	2.5219	0.5569	0.8439	0.9763	1.0153
Average total time spent carrying out search task	14.5663	8.4280	8.1992	10.0174	8.0034

Table 6.3: Task completion time, in seconds, split by interface.

Interface	Img	Tag	Thum	VSnip
Tag	0.9256	-	-	-
Thum	0.9533	1	-	-
VSnip	0.8953	1	0.9997	-
Txt	0.2495	**0.0371**	**0.0492**	**0.0289**

Table 6.4: Results of Tukey's HSD test for total time required to answer the search topics.

shows that a slight improvement was generally observed on the time required to select an answer when additional visual summaries were presented; however, the differences were not significant ($F = 1.7904, p = 0.1313$).

Table 6.3 also shows the average total time, in seconds, that was spent on carrying out each given search task, for each of the five interfaces. An F-test shows significant differences ($F = 3.0638, p = 0.0173$). Multiple comparisons were analysed using Tukey's HSD test. Users were able to finish answering the search topics with additional thumbnail, tags or visual snippet instruction compared to the text-only interface. These differences are significant according to Tukeys HSD test, as shown in Table 6.4. The results also suggest that salient images are not as helpful as the other visual interfaces for navigational search topics. We also, measured the time required to first selection see Table 6.3, and the analysis of this shows same significant differences as shown on the analysis of the total time.

We also studied the amount of time taken from selecting the answer to the time that

Interface	Img	Tag	Thum	VSnip
Tag	0.9984	-	-	-
Thum	0.8871	0.9696	-	-
VSnip	1	0.9956	0.8502	-
Txt	**0.0068**	**0.0024**	**0.0002**	**0.0090**

Table 6.5: The results of Tukey's HSD test for the time required from selection to the end of search session split by interface.

the user clicked on the "Finish" button. Users might spend this time double checking the selected answer, comparing the selected answer with other result items, or continuing to check the rest of the search results list. An F-test shows significant differences for this period of time ($F = 5.7657, p = 0.0002$), and a follow up using Tukey's HSD test was conducted as shown in Table 6.5. By comparing the visual interfaces with text-only interface, users required significantly less time (on average over 1.5 seconds less) after selecting an answer to the end of the task with the visual interfaces: Img ($p = 0.0068$), Thum ($p = 0.0002$), Tag ($p = 0.0024$) and VSnip ($p = 0.0090$). The results indicate that users were more confident finishing the task after selecting an initial answer when additional visual summaries were presented.

6.2.3 Viewing and re-viewing search result items

We extracted the total number of uniquely viewed items and the percentage of items that were re-viewed (items that were viewed more than once) per search topic from eye tracker trails. Figure 5.4, in Chapter 5, shows the mask used for the eye tracker. The regions of the textual and visual summaries of each abstract were designed to collect both the number of views, and time spent on each item.

For the visual interfaces, an item was counted as viewed if the user viewed either the

Interface	Txt	Thum	Tag	Img	VSnip
Uniquely viewed summaries (text or visual summaries)	220	186	183	199	192
Uniquely viewed text summaries	220	167	160	182	166

Table 6.6: The total number of uniquely viewed items split by interface.

textual summary or the visual summary of that item. Table 6.6 shows the number of uniquely viewed items (out of a possible 250 result items) for each condition before an answer item was selected. In Chapter 5, we collected the uniquely viewed items to finish answering the given search topics, because with informational topics, user might select more than one answer. However, with navigational topics users were asked to select one correct answer, we collected, therefore, uniquely viewed items to selecting the answer. The difference between interfaces is not significantly different ($\chi^2, p = 0.3499$). The total number of uniquely viewed text summaries (that is, the total number of single textual items that were viewed), by each user, was collected per interface as shown in Table 6.6. The results showed a significant difference ($\chi^2, p = 0.0103$). Pair-wise tests were then used to compare the text-only interface with the four visual interfaces. Apart from the image interface ($\chi^2, p = 0.0581$), users viewed significantly fewer textual items when additional visual summaries were presented; Thum ($\chi^2, p = 0.0071$), Tag ($\chi^2, p = 0.0021$) and VSnip ($\chi^2, p = 0.0060$). This result was a substantive indication that (possibly apart from the salient image interface), users relied on visual summaries to skip reading some of the text summaries. In other words, users frequently used visual summaries to decide whether or not to read the related text summaries.

In this study, the total number of times that users viewed search result items, whether textual or visual summaries or both, was counted to evaluate re-viewing behaviour. The

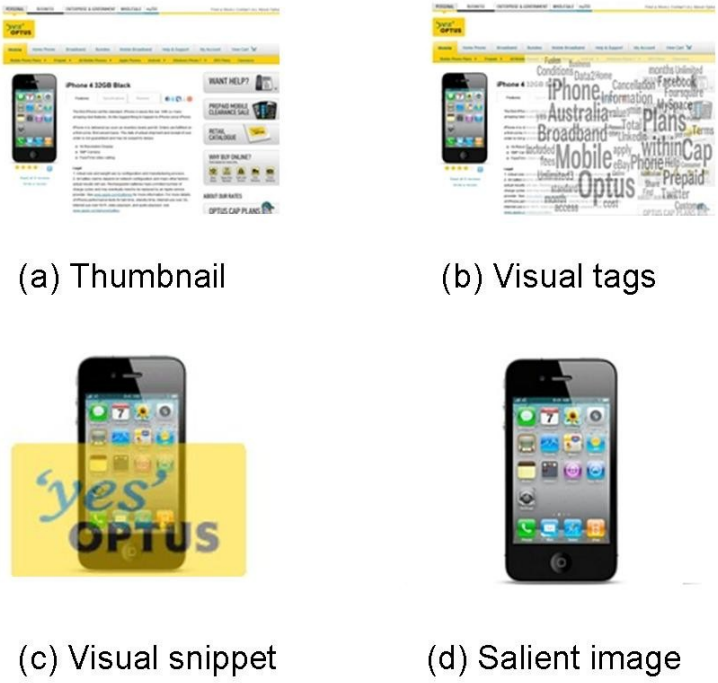

Figure 6.1: Examples of the four types of visual summaries.

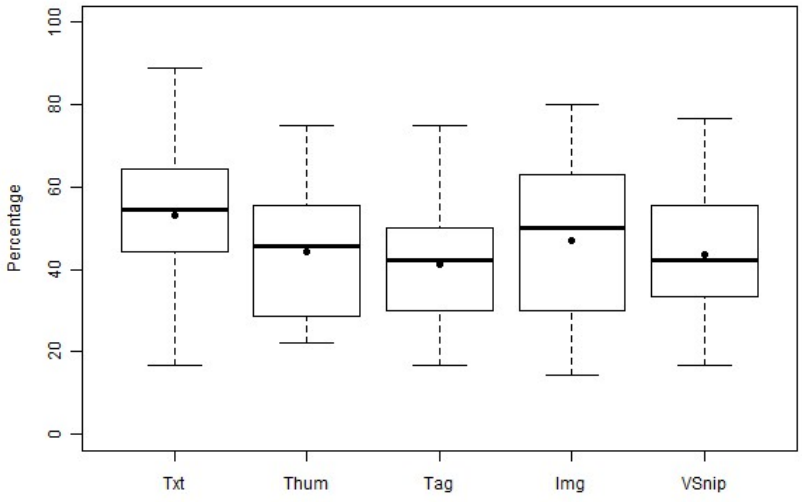

Figure 6.2: The percentage of re-viewing search result items for the entire time required to answer search topics split by interface.

Interface	Img	Tag	Thum	VSnip
Tag	0.3073	-	-	-
Thum	0.9137	0.8211	-	-
VSnip	0.8171	0.9169	0.9994	-
Txt	0.3038	**0.0014**	**0.0454**	**0.0238**

Table 6.7: The results of Tukey's HSD test for pairwise comparison of the percentage of re-viewing for the entire time required on answering the search topics.

following formula was used to calculate the percentage of re-viewing of search result items in each session:

$$\text{Re-viewing} = \frac{Total\ viewed - Uniquely\ viewed}{Total\ viewed}$$

Figure 6.2 shows the percentage of re-viewing of search result items for each session, (that is the entire time required to answer the search topics). An F-test shows a significant difference ($F = 4.2919, p = 0.0023$), and a follow up Tukey's HSD test shows significant differences as summarised in Table 6.7. The results show that, apart from the salient image interface, users required less effort (re-viewed fewer search items) when additional visual summaries were presented.

6.2.4 Interaction with text summaries

Data captured using an eye tracker can provide rich information about user attention devoted to different areas of the screen. In this study, we investigated how the user interacted with text summaries when additional visual summaries were presented, and compared this with the text-only interface. Figure 6.3 shows the time that each user spent viewing text summaries for the five interfaces. An F-test shows that there are statistical differences in the time spent looking at text summaries between the five interfaces ($F = 3.9170, p = 0.0040$). Follow up tests (Tukey's HSD) were conducted to evaluate pairwise differences among the means,

Interface	Img	Tag	Thum	VSnip
Tag	0.7048	-	-	-
Thum	0.8000	0.9999	-	-
VSnip	0.8639	0.9986	1	-
Txt	0.2306	**0.0079**	**0.0136**	**0.0217**

Table 6.8: The results of Tukey's HSD test for the time spent on text summaries split by interface.

as shown in Table 6.8. The results show that compared with the text-only interface, users spent significantly less time looking at text summaries when using Thum ($p = 0.0136$), Tag ($p = 0.0079$) and VSnip ($p = 0.0217$) interfaces. In other words, users read less text when these additional visual pieces of information are available. This result was expected as the additional visual summaries distract the users attention from the text summaries. However this significant difference could also indicate that a user spends more effort extracting the presented information with text-only interface, particularly for navigational search topics.

The next section investigates this behaviour further, thereby providing deeper insight into the impact of additional visual summaries on text summaries.

Scan-paths

An eye tracker gathers two main measurements to represent the point of a users gaze on the screen: fixations (the deliberate time required to view an object); and saccades (quick movements between fixations). In this study, a further analysis of the various effects of visual summaries on user seeking strategies, particularly on text summaries was conducted by considering scan-path, a completed observed path of eye movement sequences (fixations and saccades) across a screen. A scan-path can provide valuable insights into information seeking behaviours and cognitive style including user mental effort.

Scan-paths can be measured in multiple ways: in this study we use the scan-path length, duration and the total number of scan-path visits to a region, to measure the similarity between two paths. Scan-paths are calculated for each area of interest (AOI), where each textual or visual summary was defined as an AOI, as shown in Figure 5.4, in Chapter 5. The start of each scan-path is determined by the moment at which the user starts to look at the defined AOI. The end of the scan-path is the point at which the user leaves that AOI. An example is provided in Figure 6.4. This also allows us to count the number of times that the user views a specific AOI, which corresponds to the total number of scan-paths. The scan-path duration is the total time of gaze on a specified sequence of saccades and fixations.

Scan-paths consist of gaze samples and each gaze sample has coordinates on the screen, so that counting the distances between gaze samples will provide us with the scan-path length. Scan-path length is measured in pixels. Euclidean distance is used to calculate the distance between the coordinates for each pair of consecutive gaze samples (x_i, y_i) and (x_{i+1}, y_{i+1}) [Rao et al., 2002; Hart et al., 2009]. Three parameters were collected per session for this analysis: the total number of scan-paths; the average duration of scan-paths; and the length of scan-paths. In this analysis, only the textual scan-paths (scan-paths occurring on text summaries) were used to find the impact of presenting additional visual summaries on each search result page. Our hypothesis is that additional visual summaries can improve the users ability to judge the information presented in the text summaries. Therefore, we expect the scan-paths on textual summaries to be longer in distance and fewer in number, where visual summaries are also presented.

Figure 6.5 shows the total number of textual scan-paths for the different interfaces. The

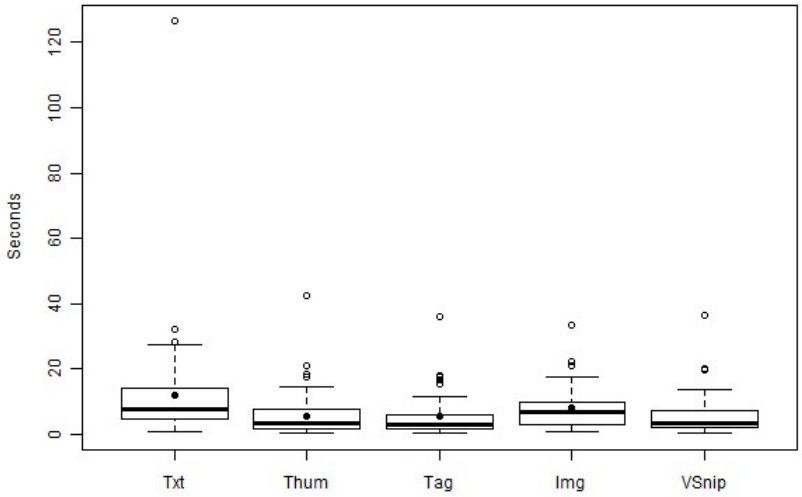

Figure 6.3: The total time spent on text summaries split by interface.

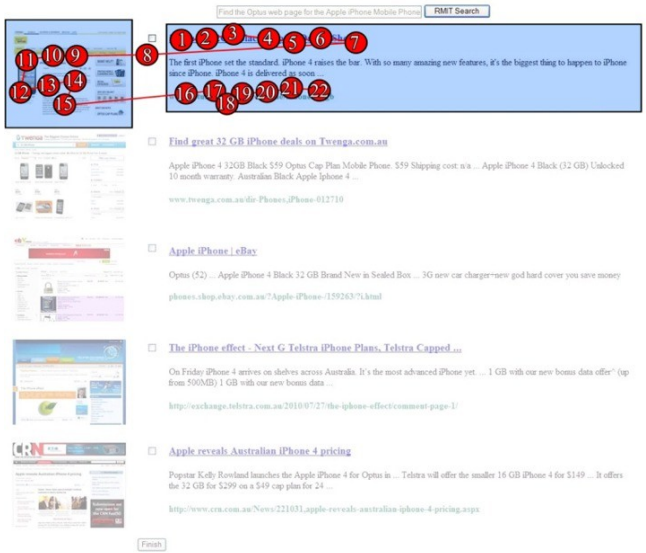

Figure 6.4: Scan-paths are gaze samples that occur in defined areas of interest (AOIs). In the image above, gaze samples 1 to 7 from one textual scan-path, and gaze samples from 16 to 22 from another textual scan-path for same text summary. The duration of the scan-path is the total duration of the gaze samples that are part of same scan-path; the length (distance) of a scan-path is calculated by the Euclidean distance between the gaze points of the path.

Interface	Img	Tag	Thum	VSnip
Tag	1	-	-	-
Thum	0.5260	1	-	-
VSnip	1	1	1	-
Txt	1	0.5260	**0.0420**	0.8750

Table 6.9: The results of Wilcoxon rank sum tests for textual scan-path length.

differences are significant ($\chi^2, p < 0.0001$). Paired follow-up tests show that users viewed text summaries significantly fewer times when any of the additional visual summaries was presented ($\chi^2, p < 0.0001$).

Textual scan-path length represents the amount of text region that users viewed. In this study, the average of the textual scan-path length was collected for each session. The results of the pairwise Wilcoxon rank sum tests across the different interfaces are shown in Table 6.9. The results show a significant difference between the text-only interface and the thumbnail interface ($p = 0.0420$). In other words, users with thumbnail interface tend to view a significantly smaller amount of text than when presented with a text-only interface.

6.2.5 Perceived search difficulty

In this study, user feedback on the perceived difficulty of each search was collected by asking participants at the end of each session to indicate how difficult they found a search on 5-point ordinal response scale:

Finding the required information for this topic was: (Very Easy / Easy / Neutral / Difficult / Very Difficult)

Figure 6.6 shows the user responses to the difficulty of finding the desired information split by interfaces. The results show a significant difference in the responses of users on the

Interface	Img	Tag	Thum	VSnip
Tag	**0.0011**	-	-	-
Thum	$P < 0.0001$	0.5152	-	-
VSnip	**0.0011**	1	0.5152	-
Txt	0.3139	**0.0235**	**0.0036**	**0.0235**

Table 6.10: The results of pair-wise comparison (χ^2) for user feedback on finding difficulty split by interfaces.

difficulty of finding the required information ($\chi^2, p = 0.0002$). Table 6.10 shows the results of pair-wise comparison showing significant differences when comparing the text-only interface with Thum ($\chi^2, p = 0.0036$), Tag ($\chi^2, p = 0.0235$) and VSnip ($\chi^2, p = 0.0235$). In other words, users found answering the given search topics with these visual interfaces significantly easier than with text-only interface. Also, a significant difference was found between the Thum interface and the other visual interfaces; Tag ($\chi^2, p = 0.0011$), Img ($\chi^2, p < 0.0001$) and VSnip ($\chi^2, p = 0.0011$). This indicates that users find the thumbnail interface significantly easier compared with the other visual interfaces. In contrast a significant difference was found when comparing Img interface with Tag and VSnip interfaces ($\chi^2, p = 0.0011$).

6.2.6 The impact of visual attention

Understanding the impact of presenting additional visual summaries on user seeking behaviour is essential for promoting comprehension and the effective use of visual summaries on web search interfaces. Therefore, in this section user seeking behaviour and performance are analysed based on the attention spent on visual summaries while answering the given search topics, which we call visual attention. Eye tracking captures the users focus on the screen, which allows the attention spent on a particular area to be calculated. In this way we are able to compare the percentage of attention paid to visual summaries with the attention

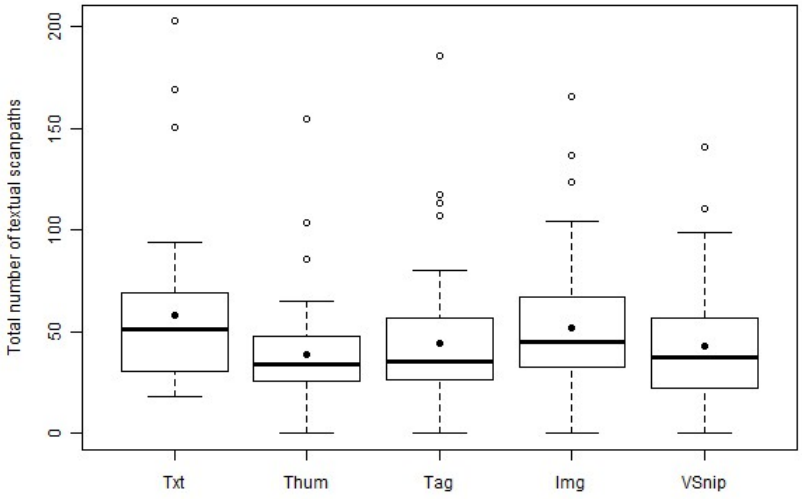

Figure 6.5: The total number of textual scan-paths split by interface.

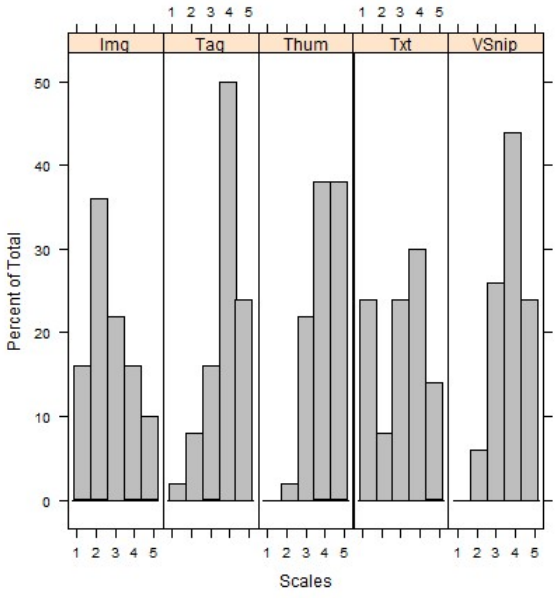

Figure 6.6: User responses to the question of how difficult it was to find the required information split by interface. On the scale 1 represents "Very Difficult" and 5 is "Very Easy".

paid to summaries overall. The visual attention is calculated for each session by dividing the total time spent on visual summaries by the total time spent on visual and text summaries. In other words, visual attention is the percentage of time spent on visual summaries relative to the total time spent on the informative components (visual and text summaries).

Visual attention distribution

Factors that might impact on visual attention include the type of visual summaries presented, the topic, or user experience. We analysed factors to obtain a richer understanding of a users information seeking strategies.

Figure 6.7 shows the percentage of user attention paid to visual summaries for each summary type. An F-test shows a significant difference in the percentage of visual attention for different interfaces ($F = 3.5770, p = 0.0149$). The results of pairwise follow-up comparisons using Tukey's HSD test are shown in Table 6.11, indicating that users paid significantly more attention to visual summaries with the VSnip ($p = 0.0153$) and Thum ($p = 0.0493$) interfaces than with the Image interface. In other words, the visual attention of users is not consistent among the four visual interfaces. User experience with the presented visual approach is interpreted by the differing amounts of visual attention. It is possible that different search topics might also lead to differences in visual attention behaviour. We therefore investigated the variation of visual attention across topics as shown in Figure 6.8. An F-test shows significant differences ($F = 2.5092, p = 0.0433$). A pairwise follow-up with Tukey's HSD test shows only one significant difference between the ARIA topic and the Facebook topic ($p = 0.0349$). This result suggests that the different aspects of navigational search topics, such as finding

Interface	Img	Tag	Thum
Tag	0.1652	-	-
Thum	**0.0493**	0.9537	-
VSnip	**0.0153**	0.7832	0.9754

Table 6.11: The results of Tukey's HSD test for the percentage of visual attention split by interface.

the homepage or a single particular webpage, might have an impact on user attention on the visual summaries as can be seen from the variation in the inter-quartile ranges for different topics in Figure 6.8. User responses for the perceived difficulty of a given topic were also examined in terms of visual attention. The results showed no significant differences $(F - test, p > 0.05)$.

Each user in our sample carried out five search tasks, four of which used visual interfaces. Figure 6.9 shows the percentage of attention paid to visual summaries for these 200 sessions, split by user. Users are ordered by the increasing mean of their visual attention over the four visual interfaces. As can be seen, visual attention behaviour varies markedly from person to person $(F = 3.8027, p < 0.0001)$.

As can be seen from the preceding analysis, visual attention behaviour varies with the type of interface that is presented, with the search topic that is being pursued, and with other user factors. To better understand this behaviour, and to analyse the impact that it may have on the search results, we re-analysed our data based on visual attention groups.

The 200 sessions, where users searched with the visual interfaces, were categorised based on the percentage of visual attention given by users in each session. This shows, interestingly, that in 50 of the 200 sessions, users did not look at the visual summaries at all. In these sessions, users can be described as displaying verbal behaviour: they prefer reading

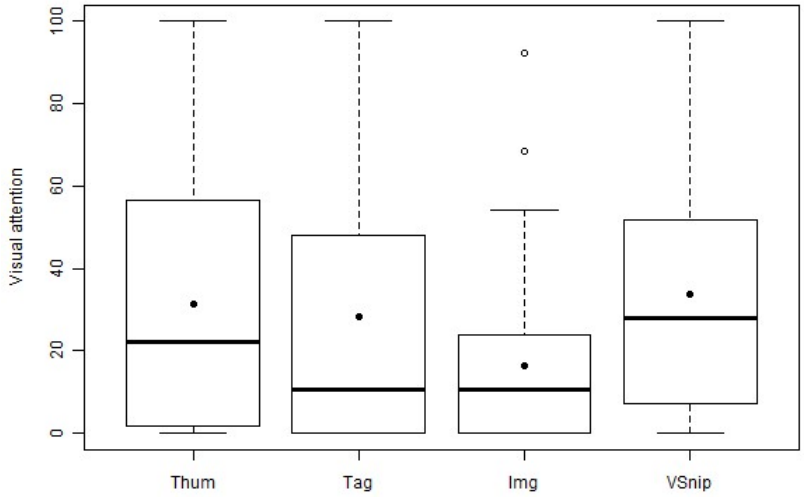

Figure 6.7: The percentage of visual attention of users split by interface.

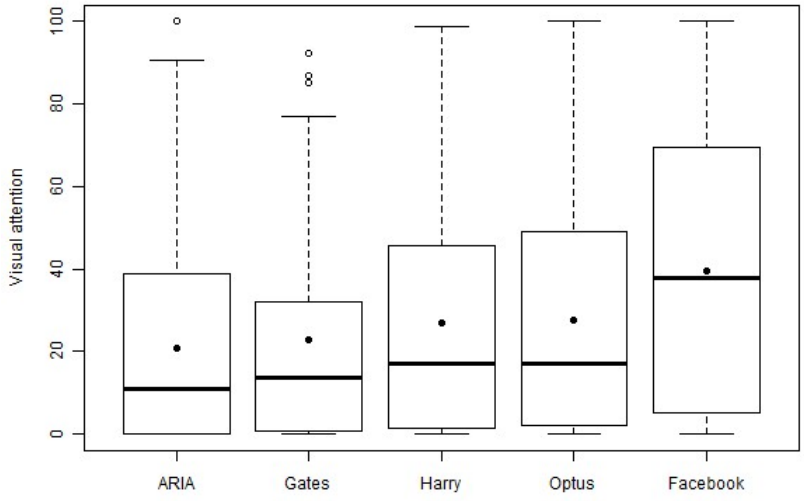

Figure 6.8: Visual attention split by topics.

Number of sessions related to the same user	One	Two	Three	Four
Users	10	9	6	1

Table 6.12: The number of users in the 50 non-visual sessions split by frequently of sessions.

text summaries rather than viewing visual summaries. In other words, these users with a particular interface and a specific type of topic prefer to rely only on viewing text summaries to predict the relevant answer. Further analysis showed that these 50 sessions comprised data from only 26 users. Table 6.12 shows the number of sessions occurring in the 50 non-visual sessions that are from the same user. This result shows that user behaviour can change from visual (attention substantially paid to visual summaries) to verbal (attention paid rapidly to text summaries) among the given search topics and interfaces. This could indicate user experience, preference, or the presenting approach for visual summaries. Interestingly, one user did not look at the visual components at all.

To continue our analysis of the impact of visual attention, we divided the 200 sessions according to the percentage of visual attention into four equal-side categories: non-visual, light, mid and heavy: each category therefore consisted of 50 sessions. The aim was to understand the impact of visual attention on browsing behaviour with regard to the informative components of the web search result page. In this analysis, we examined user behaviour and performance when the user was non-visual (that is, had no interest in browsing visual summaries even when such visual summaries were offered as part of the interface).

In the subsequent sections, we re-analysed the search outcomes effectiveness of relevance prediction, task completion time, re-viewing behaviour, interaction with summaries, scanpaths, and perceived difficulty but in terms of visual attention behaviour, rather than by

Category	Non-visual	Light	Mid	Heavy
Correct	28	37	42	43
Incorrect	22	13	8	7

Table 6.13: The total number of correct and incorrect answers selected by users for the four visual interfaces, split by category.

interface.

Effectiveness of relevance prediction

Recall that in the user study that was conducted, participants were asked to select one relevant answer for each given search topic. Table 6.13 shows the total number of correct and incorrect choices that were made by users, split by the previously defined visual attention behaviour categories. The results show that users, in heavy visual sessions, correctly selected more relevant answers compared those with fewer visual sessions. A χ^2 test on differences in the total number of incorrect answer shows a significant difference ($\chi^2, p = 0.0103$). Pairwise following between the four categories showed significant differences between the non-visual sessions compared with the heavy visual sessions ($p = 0.0053$) and mid-visual sessions ($p = 0.0106$). Users who paid more attention to visual summaries were significantly more successful in predicting relevant answers.

Task completion time

Figure 6.10 shows that the total time spent in answering the given search topics, in the four visual behaviour categories: non-visual, light, mid and heavy. An F-test shows significant differences ($F = 6.8765, p = 0.0002$). Multiple comparisons were analysed using Tukey's HSD test as shown in Table 6.14. Interestingly, apart from light visual sessions ($p = 0.0062$),

Category	Heavy	Mid	Light
Mid	**0.0491**	-	-
Light	**0.0003**	0.3977	-
Non-visual	0.8228	0.3096	**0.0062**

Table 6.14: The results of Tukey's HSD test for total time spent in answering the given search topics split by the four visual categories.

the results show no significant difference between non-visual sessions and the other session categories (mid ($p = 0.3096$) as well as heavy visual sessions ($p = 0.8228$)). Users in sessions with heavy visual attention were able to finish answering the search topics in significantly less time (on average 4 to 6 seconds quicker) than those with light ($p = 0.0003$) and mid ($p = 0.0491$) visual sessions.

Viewing and re-viewing search result items

The number of uniquely viewed items out of the possible five search results was collected for each session; an item was counted as viewed if the user viewed either the textual summary or the visual summary of that item. Table 6.15 shows the number of uniquely viewed items for the four visual behaviour categories. The results show a significant difference ($\chi^2, p = 0.0019$). Paire-wise comparisons show a significant difference for the number of uniquely viewed items between non-visual sessions compared with light ($\chi^2, p = 0.0002$), mid ($\chi^2, p = 0.0024$) and heavy ($\chi^2, p = 0.0291$). This is not surprising due to the way in which the categories are constructed.

Since the four categories were divided based on visual attention, users with non-visual behaviour were expected to spend more time on text summaries than users in heavy visual categories. However, the total number of uniquely textually viewed items varied between the

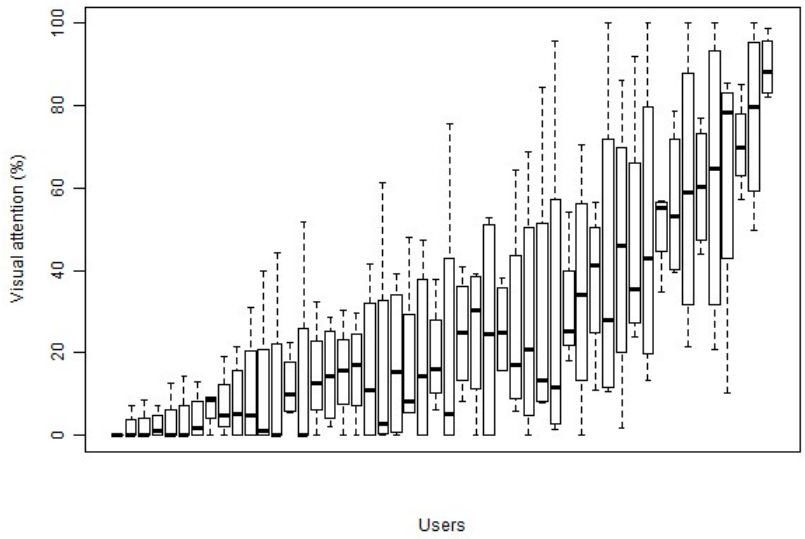

Figure 6.9: The percentage of visual attention ordered by the mean visual attention across four visual interfaces.

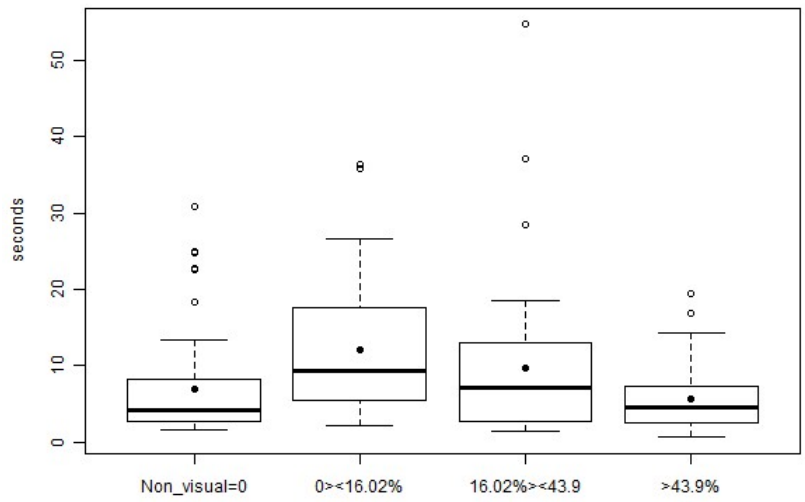

Figure 6.10: The total time spent in answering the search topic for the four visual interfaces split by the four visual categories.

category	Non-visual	Light	Mid	Heavy
Overall uniquely viewed items (incl. visual summaries)	148	219	205	188
Uniquely textual viewed items	148	180	112	84

Table 6.15: The total number of uniquely viewed items to select answers split by visual categories.

Category	Heavy	Mid	Light
Mid	**0.0455**	-	-
Light	$P < 0.0001$	**0.0256**	-
Non-visual	$p < 0.0001$	**0.0256**	0.0769

Table 6.16: The results of the statistical pair-wise tests (χ^2) between the four categories.

four categories. The total number of uniquely viewed textual items was collected per interface as shown in Table 6.15. The results showed a significant difference ($\chi^2, p < 0.0001$). Pair-wise tests were used to compare the four categories, the results being shown in Table 6.16. The results show that, apart from light visual sessions ($\chi^2, p = 0.0769$) users with visual attention viewed significantly fewer text summaries compared with non-visual users. In other words, users with high visual attention (mid and heavy) relied on visual summaries to skip viewing some text summaries.

The proportion of re-viewed items was calculated for selecting answers so as to finish answering the given search topics for each session, as described in Section 6.2.6. An F-test shows no significant difference between the four visual categories neither for selecting answers ($F = 2.3336, p = 0.0752$), nor for finishing the answering of each task ($F = 1.0187, p = 0.3855$).

Interaction with text summaries

The averages of length, duration and total number of textual scan-paths (sequence of fixation and saccades occurring on text) were collected for each session as explained in Section 6.2.4. The four visual categories presented a different range of behaviours on the amount of visual attention spent by a user in a session.

Studies show a correlation between user attention and concentration, such as paying more attention to interesting information [Shimoda, 1993], or important information [Flammer and Kintsch, 1982]. Furthermore, Rayner [2009] found that a user requires more time in scene perception than in reading. Therefore, we evaluated the average number of textual scan-paths before selecting answers so as to examine the impact of the amount of visual attention spent by users on the concentration of users in reading the text summaries. Figure 6.11 shows the average number of textual scan-paths split by visual behaviour categories. An F-test shows a significant difference in the average number of textual scan-paths ($F = 10.8791, p < 0.0001$); a follow up of a pair-wise Wilcoxon signed rank test shows that users with heavy visual attention spent less time viewing text summaries than with the other three categories: non-visual ($p = 0.0388$), light ($p < 0.0001$) and mid ($p = 0.0063$) as shown in Table 6.17. In addition, the results show that users with light visual attention significantly concentrate more on reading text than non-visual users ($p = 0.0008$). This contrasts with non-visual users when compared with mid-visual attention users who show no significant difference ($p = 0.3242$).

The average duration and length of textual scan-paths for selecting answers were collected for each session to examine user speed in reading text summaries. User speed in reading text summaries was measured by dividing the average textual scan-path length by the average

Category	Heavy	Mid	Light
Mid	**0.0063**	-	-
Light	$p < 0.0001$	**0.0241**	-
Non-visual	**0.0388**	0.3242	**0.0008**

Table 6.17: The results of a pair-wise Wilcoxon signed rank test for the average number of textual scan-paths when selecting an answer, split by visual categories.

Category	Heavy	Mid	Light
Mid	0.0768	-	-
Light	**0.0332**	0.9880	-
Non-visual	**0.0156**	0.9352	0.9936

Table 6.18: The results of Tukey's HSD test for user's speed in reading text summaries to selecting answers split by visual categories.

duration of textual scan-path. The results are shown in Figure 6.12. An F-test shows significant difference ($F = 3.8270, p = 0.0108$). The results of Tukey's HSD test are shown in Table 6.18; apart from mid-visual sessions ($p = 0.0768$), a user with heavy visual attention read text summaries significantly faster than non-visual ($p = 0.0156$) and light visual ($p = 0.0332$) users. This result supports the hypothesis that users with heavy visual attention tend to skim text summaries rather than read the text, in contrast to users with lower visual attention who tend to read text summaries in a similar manner to non-visual users.

Search topic difficulty

In this study, user feedback on the perceived difficulty of each search topic was collected as described in Section 6.2.5. Results show no significant difference between the four visual categories on the responses to the difficulty ($\chi^2, p = 0.1194$). This result demonstrates that in this study, the amount of attention spent on visual summaries is not influenced by any perceived difficulty in the search topics.

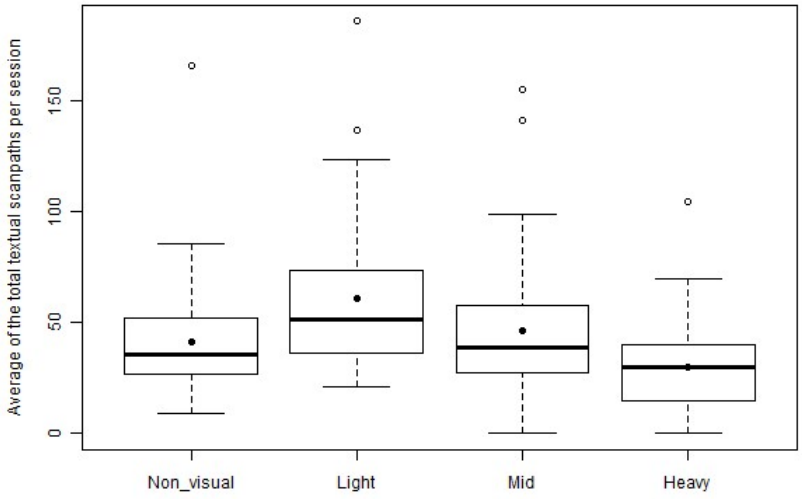

Figure 6.11: The average number of textual scan-paths split by visual categories.

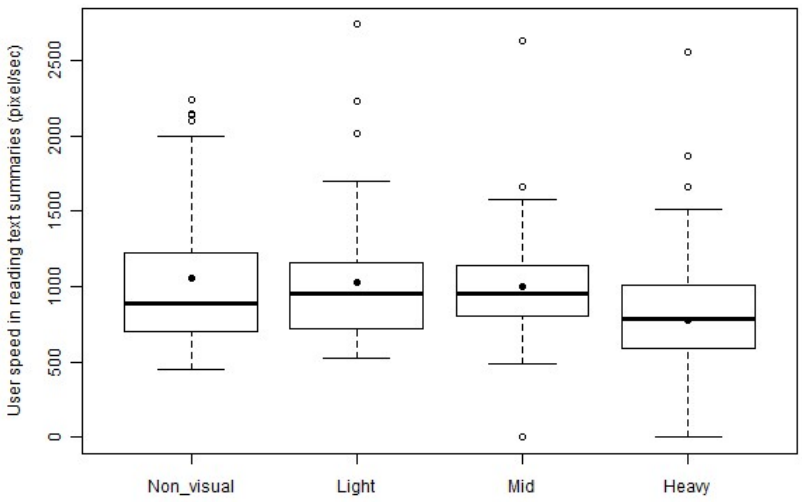

Figure 6.12: User speed in reading text summaries for selecting answers split by visual categories.

6.2.7 Forms of search results viewing behaviour on visual interfaces

This study investigates the effect that the presence of additional visual summaries can have on user browsing behaviour. Previous sections have considered higher level aspects of search, such as attention devoted to aggregated areas of interest. We now focus on more fine-grained views of searching behaviour. Diagrammatic representations were created to get a firmer view of user interaction with the search results when additional visual summaries were presented. This analysis involved the 200 sessions of the four visual interfaces. This section studies user behaviour up to answer selection time only, as this is a critical period for processing, assessing and interpreting evidence to predict the relevant answer.

Examples of the interaction diagrams that were created to study the detailed user strategies when browsing visual interfaces are shown in Figure 6.13. The diagram plots the users eye movement across the screen for the entire period of time required to answer the given search task. The vertical axis represents the rank position of the answer components, with the textual abstract items shown above the x-axis, at the top and the visual summary shown below the x-axis. The horizontal axis represents the time spent, in seconds, answering the task. The solid line in the diagram illustrates the users eye movement; horizontal lines at a particular level indicate a period of attention where the gaze is focused on that item. Line segments on the horizontal axis indicate that the user was looking at white space (that is, their gaze was outside a defined area of interest). The selecting action, where the user clicked on their chosen answer item, is shown by a vertical dashed line.

The 200 user browsing diagrams were initially analysed based upon the amount of attention spent on visual summaries as described in earlier sections. Then sequences of user

Interface	Non-Visual	Light	Mid	Heavy
Thum	12	9	14	15
Tag	18	10	8	14
Img	14	18	13	5
Vsnip	6	13	15	16

Table 6.19: The total number of interfaces in each visual category.

viewing behaviour were evaluated and split into four visual behaviour categories: heavy, mid, light and non-visual, as described in Section 6.2.6. Table 6.19 shows the total number of each category that occurred for each interface. Further analysis of user actions when viewing the search results showed that each of the four groups has subcategories that demonstrate consistent characteristics such that particular behaviour browsing forms can be identified. Examples are shown in Figure 6.13.

Heavy visual: 50 sessions were identified as extensively visual, where users spent more than 40% of their answering time looking at the visual region of the interface, and only one or two textual items were viewed. In heavy visual sessions, two common strategies were identified when browsing the search results. In most of the heavy visual sessions (66% of the heavy visual sessions), users viewed only visual summaries or read only a single textual summary, which was typically the selected answer. In the other heavy visual sessions (44%), as well as looking at the additional visual summaries, users read two text summaries so as to make a decision when selecting their predicted answer.

Mid-visual: In this category (50 sessions), users browsed both visual and text summaries. Here, less attention was paid to visual summaries than in the heavy visual category , less than 40% and more than 16% of the answering time , and greater attention was paid to text summaries, with at least two different textual items being viewed. Based on the

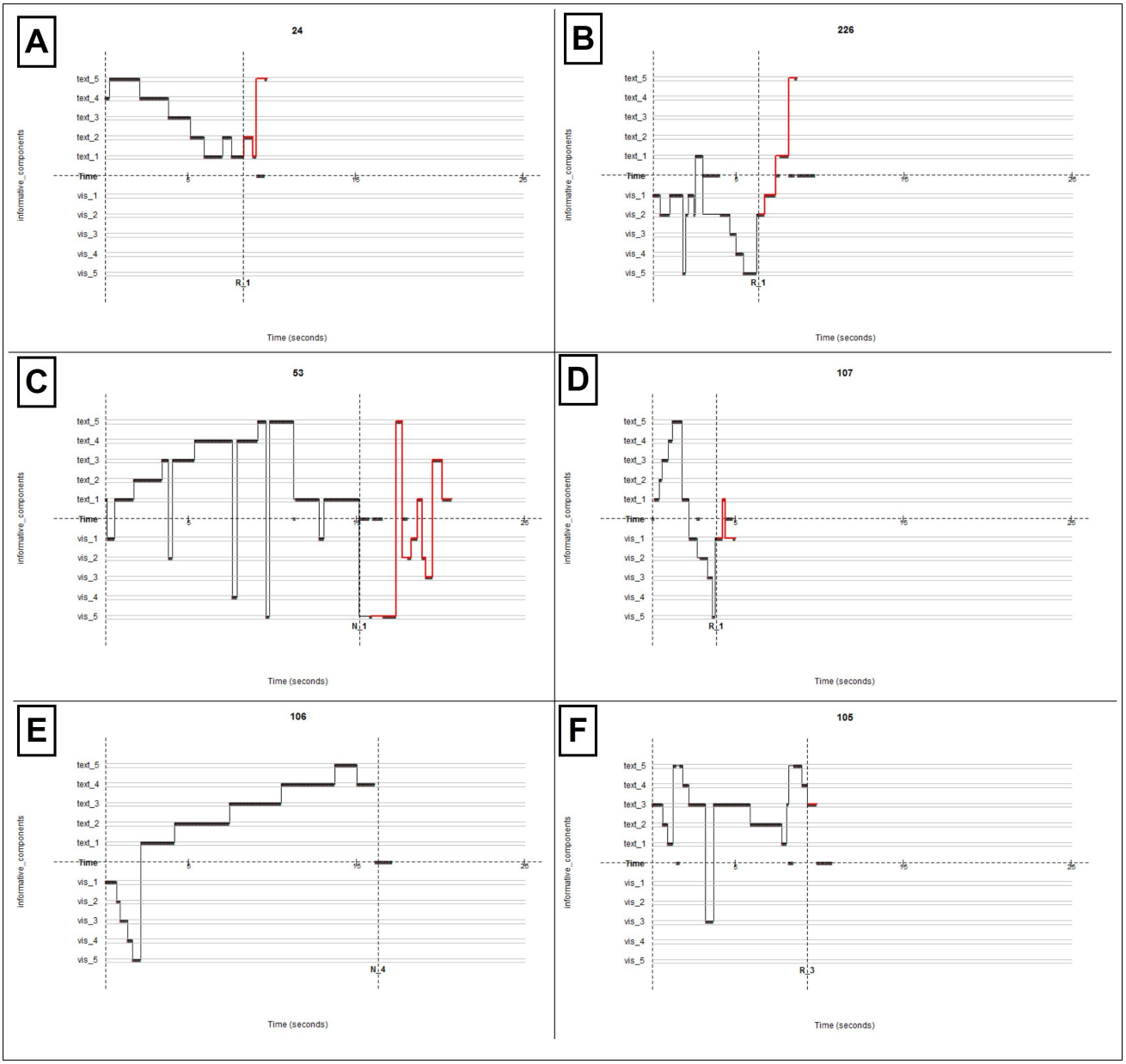

Figure 6.13: Examples for the user diagrams used to study user browsing forms: (A) Non-visual. (B) Extensive visual. (C) Neutral pairs. (D) Start by text then switch to visual summaries. (E) Start by visual then switch to text summaries. (F) Barely visual.

sequences of a users actions, three browsing forms were identified overall in the mid-visual sessions. In paired mid-visual (42% of mid sessions) user's viewed visual and text summaries in alternating sequence; that is users read the textual summary of an item and then viewed its corresponding visual summary, or vice versa. In 28% of the mid-visual sessions, users started browsing only the search results by first viewing the text summaries, viewing two or three text summaries, then switching attention to visual summaries. In contrast, in 30% of the mid sessions, users started browsing search results by first viewing visual summaries, viewing two or three visual summaries, and then switching attention substantially to text summaries, with scant attention paid to visual summaries after the toggle point.

Light visual: Users in this classification spent only 16% or less of their answering time viewing visual summaries. In 36% of light visual sessions, users viewed only one or two visual summaries, usually at the beginning of the answering time of the task, where the viewed visual summaries were mostly not selected by the user as a relevant answer. Interestingly, in the rest of the light sessions (64%), users viewed one or two visual summaries after re-viewing the text, normally just before selecting an answer for which the visual summary had already viewed.

Non-visual sessions: users in 50 sessions focused only on browsing text summaries and did not view any visual summaries at all before selecting an answer. This browsing form occurred on four visual interfaces. Interestingly, as shown in Table 6.19, the visual snippets interface had the smallest number of non-visual sessions, compared with the other three visual interfaces.

User browsing forms over all the four visual interfaces

When analysing the relationship between user browsing forms and the four visual interfaces, the results show that users can be classified into four groups: constant, pairs, sharp and unstructured. Figure 6.14 summarises the visual attention of users across each of the four interfaces, with each line representing one user across the four interface dimensions. The point on each axis represents the percentage of time spent on the visual interface. In the constant group, Figure 6.14 (A), only six users, out of the 50 users, follow the same visual browsing behaviour over all the four visual interfaces. Five users were intensive visual and the sixth was non-visual over all the four visual interfaces. The remaining 44 users follow different visual browsing behaviour over the four visual interfaces. In the pairs group, 15 users follow only two visual behaviour forms over the four visual interfaces. In most of the cases, the two forms are close to each other in terms of the visual attention behaviour, such as non-visual and low, or mid and heavy, see Figure 6.14 (D). The increment on visual behaviour was noticed for thumbnail, tag or visual snippet. The sharp group, of 16 users, consists of user who show consistent visual behaviour forms over three visual interfaces, but suddenly sharply change on the fourth visual interface, see Figure 6.14 (C). The sharp change in visual attention is noted to be an increase or drop. The sharp increase cases occurred on the thumbnail and visual snippet interfaces; in contrast the sharp drop was noted for the image interface. Finally, the unstructured group, consisting of 13 users, is where a user follows at least three different browsing forms over the four visual interfaces, see Figure 6.14 (B). No specific behaviour was recognised in this group. This suggests that presenting different types of visual summaries can impact on user behaviour. In addition, other factors might have an

impact on a particular behaviour such as user experience or interest [Flammer and Kintsch, 1982; Shimoda, 1993; Rayner, 2009].

6.3 Discussion and summary

In this study, the impact of different visual summaries on user seeking behaviour and search performance was evaluated. A series of five navigational search topics were answered using five different interfaces by fifty participants.

6.3.1 The effectiveness of different approaches for visual summaries

Results show that users with the text-only interface required significantly more time in total to finish answering the task that when provided with thumbnail, tag or visual snippet interface. Also, users with the text-only interface required (apart from the salient image interface) significantly more time after selecting an answer to complete the task; an indication of the difficulty of finding the required answer. Furthermore, user feedback about perceived task difficulty shows that users with the text-only interface experience significantly greater difficulty in finding required answers, compared with using visual interfaces. Overall, the thumbnail, visual tag and visual snippets improve user ability to predict relevant answers in significantly less time and with less effort compared with text-only interface.

Our results indicate that the salient image interface is less effective for finding relevant answers for navigational search topics. Compared with the previous experiment, in Chapter 4, where the same range of interfaces were examined with informational search topics, the salient image interface achieved better performance than thumbnail, visual tag and visual

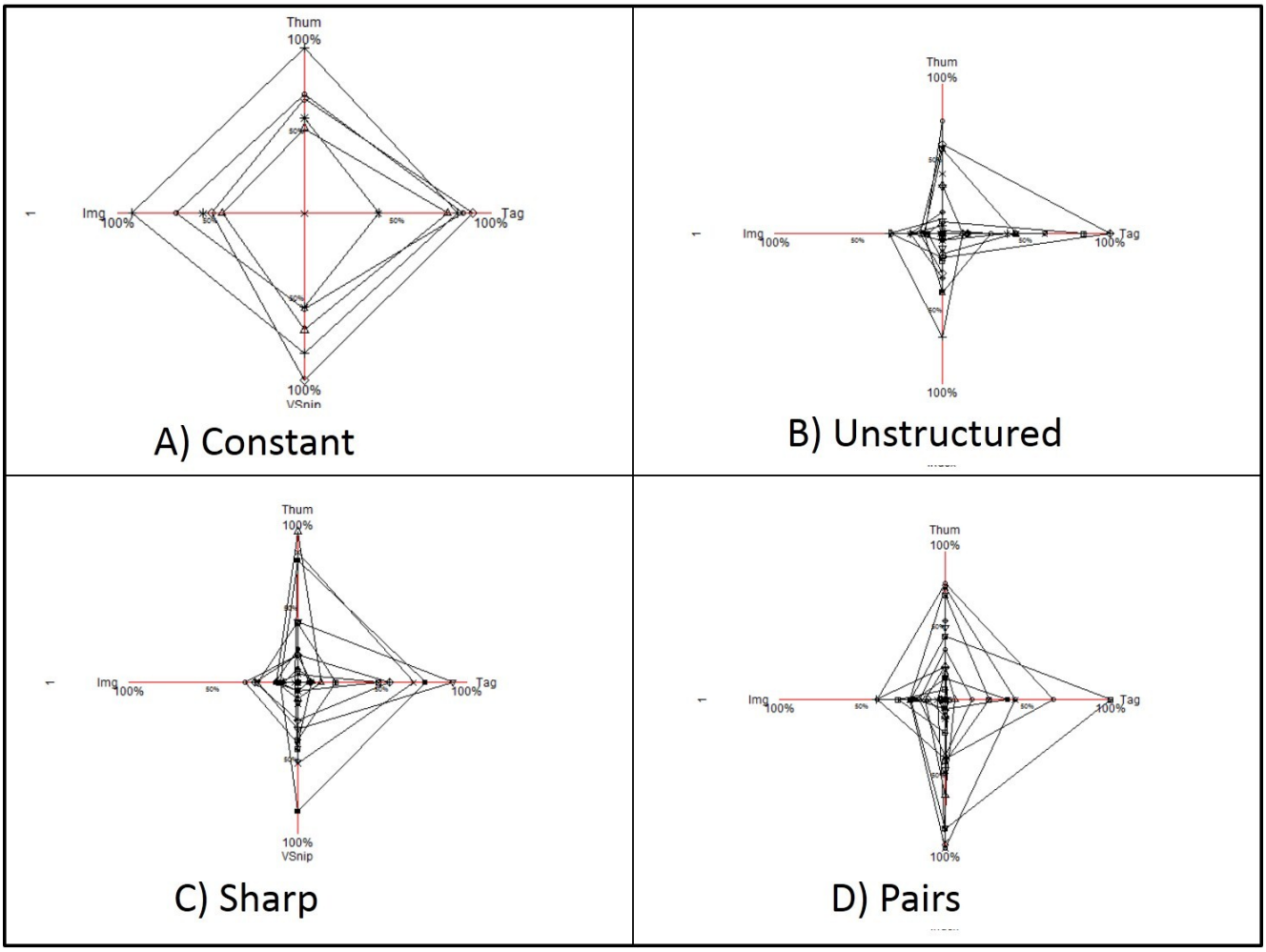

Figure 6.14: Classification of users browsing forms over all the four visual interfaces: A) Constant. B) Unstructured. C) Sharp. D) Pairs.

snippets. Thus, the type of search task (informational versus navigational) has a substantial impact on the effectiveness of different types of visual summaries.

6.3.2 The impact of visual summaries on user behaviour

User behaviour was evaluated by the percentage of the time spent on visual summaries, where the 200 sessions of visual interfaces were grouped into four categories: non-visual, light, mid and heavy. In this analysis, we were interested in evaluating the impact of different amounts of attention spent on visual summaries on user seeking behaviour. Heavy and mid-visual sessions provided the most accurate prediction of result relevance, compared with non-visual and light sessions. Presenting additional visual summaries significantly influenced users to skip reading irrelevant text summaries, by viewing instead their related visual summaries. This is underlined by the results that show no significant difference in the total time spent in identifying answers to the given search topics between non-visual and the other three visual categories. However, a significant difference was found in the total number of uniquely viewed items which showed light, mid and heavy visual users viewing significantly more unique items, with fewer uniquely viewed text summaries. The results show that paying more attention to visual summaries significantly assists in finding relevant answers.

We also investigated user strategies for browsing web search results when additional visual summaries were presented. The results show various user seeking strategies are used, even by the same user on different interfaces. A total of eight browsing forms was found across four of the categories of visual sessions: non-visual, light, mid and heavy. The results also show that user experience and familiarity with the presented visual summary might have a

substantial impact on user seeking behaviour. An example of this impact on user seeking behaviour is the sharp increase shown by the same user in attention to visual summaries from one particular approach of visual summary compared to another. Also, the different patterns of visual browsing behaviour of users suggest that presenting different types of visual summary can impact on user behaviour.

6.3.3 User cognitive processes

The key idea of presenting a visual summary is to help users in the process of making a decision to find a relevant answer. This process consists of four stages: learning, solving problems, memory and comprehension. These mental processes are called cognitive processes or information processing. The relation between cognitive process and some aspects of user searching behaviour is discussed in Section 3.2.

Users with a thumbnail interface view text summaries significantly faster than those with text-only interface. Apart from the image interface, the results outlined significant differences between text-only interface and the visual interfaces on the number of uniquely viewed items, re-viewing percentages and the number of times users viewed text summaries. These measures are indicators of the level of user cognitive effort [Goldberg and Kotval, 1999; Cole et al., 2011a; Gwizdka and Spence, 2006; Pan et al., 2004], which suggests that users required significantly greater cognitive effort to use the text-only interface to answer the given search topics. Furthermore, this result suggests that users with the text-only interface read text summaries for comprehension of the text [Masson, 1982; Reichle et al., 2006; Cole et al., 2011b]. In contrast, the results suggest that by presenting additional visual summaries

users require less mental effort to comprehend text summaries where the visual cues might help users decide whether to skip text summaries or skim text summaries instead of reading or reading just part of the textual summary.

The impact of visual summaries on user cognitive abilities based on the attention spent on visual summaries (non-visual, light, mid and heavy) can be studied based on the earlier findings of cognitive processes. Users who paid significant attention to visual summaries view text summaries significantly faster; in other words, users tended to skim text summaries so as to get a general idea of the content. In contrast, users in non-visual sessions tended to read text to process comprehension.

6.3.4 Summary

In this study, we evaluated the impact of different visual summaries on user seeking behaviour and performance. Fifty participants carried out a series of five navigational search topics using different interfaces. Our user study focused mainly on evaluating the ability of users to predict relevant answers for the given navigational search topics, and to evaluate user seeking behaviour and performance when additional visual summaries were presented.

Our analysis shows that the type of visual summary had a significant impact on user performance and behaviour for navigational search topics. Some approaches to visual summaries not only significantly improved the ability of users to find answers in shorter amounts of time, but also significantly reduced the amount of effort required to extract the information from the search result page. Also, the results demonstrate that in general users can find relevant answers to navigational search queries much more easily where visual summaries

are presented than with text-only interface. Different amounts of attention spent on visual summaries show different forms of browsing for the search results page. In future work, we plan to conduct a further user study with a wider range of search topics to compare the effectiveness of visual summaries.

Chapter 7

Comparing navigational and informational searching

According to our previous findings, topic types may have a substantial impact on user seeking behaviour and the effectiveness of presented visual summaries. Therefore, based on our findings, we identified the best-performing interface (thumbnail) for navigational topics and the best-performing interface (salient image) for informational topics. This chapter aims to address our third research question: How does the type of search topic influence the effectiveness of additional visual summaries for the presentation of web search results?

The following sub-questions are investigated:

1. To what extent do topic types impact on the effectiveness of thumbnail and image summaries, particularly in regard to users' ability to predict relevant answers, effort expended and task completion time?

2. To what extent do topic types impact on user searching behaviour, particularly on the relative attention that users pay to the informative components?

The chapter is organized as follows: in Section 7.1 the experimental framework is presented, while the results are described in Section 7.2. The discussion and summary are presented in Section 7.3.

7.1 Experimental framework

Our user study considers two interfaces (image and thumbnail), and 24 search topics (12 navigational and 12 informational). Compared with the previous user studies in this thesis, we employed a larger set of topics (5 topics in the previous studies), using various features as will be described later in this chapter. In addition, in this study we focus more on the correlation between topic types and visual summaries on user searching behaviour, and hence employ only two interfaces. At the component level, user attention to specific informative components was evaluated by collecting the amount of time that the user's gaze rested on each component. The gaze regions were closely bounded on the interface component, leaving regions of white-space between them, as shown in Figure 7.1. (In the previous user studies, we did not evaluate user gaze distribution at the component level.)

The purpose of this study is to investigate whether and how additional thumbnail and image summaries may impact on users searching behaviour. Two query types (informational and navigational) were therefore used to examine the relation between those visual summaries and user searching behavior and performance. A range of different techniques was employed to evaluate user performance, such as the time required to find the desired information, the

Figure 7.1: The mask used to collect time spent on the specific informative components (A) Exact visual summary. (B) Page title. (C) Text snippet. (D) URL.

total number of relevant selected answers, and the time to first selection. Additionally, the study evaluates the relative attention that users pay to different informative components.

7.1.1 Interfaces

We use the same template (described in Chapter 4) to design two interfaces for this experiment (thumbnail and salient image), each presenting exactly the same text summary but with different visual summaries. The template enable us to analyse the user' gaze distribution at level component, as shown in Figure 7.1. For each item, the interfaces present: a document title; a text snippet (a short text extract from the source document that closely relates to the query terms); the URL; and a visual summary component.

7.1.2 Experimental setup

56 subjects were recruited to take part in the user study, however, due to the importance of calibration and quality of recording, see Section 3.5.2, we eliminated users with bad calibration or less than 80% quality recording. The data collected from 48 subjects was included in the analysis of this study.

Each participant was asked to evaluate items in a search results list for a series of four search topics (2 informational and 2 navigational) using different interfaces for each topic. For each task, five answer items were shown on a single page. Participants used the mouse to select all items that they considered to be relevant to the given topic, and were not able to browse the actual web pages embedded in the text search results, relying solely on the search results page given. Users were presented with a fixed search results list for the topic and did not engage in interactive searching.

7.1.3 Topic selection

We evaluated the two interfaces using informational and navigational search topics. Twenty four search topics based on general knowledge were developed: 12 navigational and 12 informational.

To obtain realistic search engine results, we used Bing to collect the top ten search result items. Five search result items were chosen randomly from this set, after excluding Wikipedia entries (since they provide obvious answers for the experimental tasks). The different domains of topics were designed to meet participant interest and knowledge. The varying position of the answer, number of relevant answers, length of the query string and domains were taken

into account in the process of topic selection.

For the informational search topics, Average Precision (a single-value metric that takes the number and position of relevant answers into account, see Section 2.1.3) shows a good spread on position and number of relevant answers for each topic. For six topics, the Average Precision is less than 0.50 (topics from 1 to 6 in Table 7.1), while the others are equal or over 0.50 (topics from 7 to 12 in Table 7.1). Additionally, since the navigational topics usually have only one relevant answer, the position of the correct answer on the ranking list was taken into account in the process of selecting topics. For six of the navigational search topics, the correct answer is located in the first or second position on the ranked list (topics from 1 to 6 in Table 7.2), while the correct answer for the other six topics is located in the fourth and fifth position (topics from 7 to 12 in Table 7.2).

Index	Informational search topics	Query terms	Domain	Average Precision	ID
1	Find webpages that give the name of the director of the movie "The Impossible".	the impossible	Movie	0.20	impossible
2	Find webpages with information about the number of stars in the Chinese flag.	china	Geography	0.20	china
3	Find webpages that list the countries on Brazil's borders.	brazil borders	Geography	0.20	brazil
4	Find webpages that show how many legs are on a lobster.	lobster	Biology	0.33	lobster
5	Find webpages showing the effect of steroids on the human body.	steroids	Health	0.37	steroids
6	Find webpages with information about the meaning of "Billabong" in Australian English.	billabong australia english	Language	0.37	billabong
7	Find webpages showing problems leading to the sinking of the Titanic.	titanic	History	0.45	titanic
8	Find webpages showing the impact of soft drinks on your health.	soft drinks	Health	0.50	soft
9	Find webpages showing the location of the smallest penguin in Australia.	small penguin australia	Biology	0.59	penguin
10	Find webpages with information about what IP refers to in a computer network.	ip	Computer	0.64	ip
11	Find webpages showing how many countries Europe includes.	europe	Geography	0.75	europe
12	Find webpages with information about butterflies in Australia.	butterfly australia	Biology	0.76	butterfly

Table 7.1: Informational search topics.

Index	Navigational search topics	Query terms	Domain	Position of relevant answer	ID
1	Find the homepage of Fantastic Furniture.	fantastic	Shopping	1	
2	Find the Melbourne Central cinema web page on the Hoyts website.	melbourne central cinemas	location	1	hoyts
3	Find the Fox channel web page on Youtube.	youtube fox	video	1	fox
4	Find the official website for Mount Buller in summer.	mt buller	Tourist	2	buller
5	Find the official website for Rebecca Black.	rebecca black	Music	2	Black
6	Find the web page for the pink Samsung galaxy s2 on the Vodafone website.	vodafone pink samsung galaxy s2	Shopping	2	galaxy
7	Find the official website of Virgin Mobile at Australia.	virgin	Mobile	4	virgin
8	Find the official homepage of the movie "Real Steel".	real steel movie	Movie	4	steel
9	Find the eBay web page for buying a webcam camera.	laptop camera webcam	Shopping	4	webcam
10	Find the sport news web page on The Age website.	sports news	Sport	5	age
11	Find the Disney web page on the Yahoo website.	disney games	Games	5	disney
12	Find the official Sun web page for the London Olympics 2012 news.	olympic	News	5	olympic

Table 7.2: Navigational search topics.

7.1.4 Procedure

After a subject read the experiment instructions on the screen, a topic was shown. The subject clicked a start button to load the search interface. Five items were displayed as search results, and subjects were instructed to select the items that they consider to be relevant answer(s) for the task. After that, the subject clicked on a finish button to move to the next task.

The presentation of topics and interfaces was determined by a Latin square to control for topic and interface order effects. (See Section 4.1.4.)

7.2 Results

Different metrics were used to analyse the effectiveness of visual summaries and the impact of topic types on user seeking behaviour.

7.2.1 Search success

Subjects were asked to select relevant answer among the presented search result items for the given search topics. The tasks in general were framed as seeking answer(s) that meet the informational needs expressed in the topic, and subjects were allowed to mark one or more answers for both the navigational and informational topics. We did not explicitly distinguish between the two types of topics, but the wording of individual topics was implicitly different: for example, the navigational search topics were all expressed along the lines of "find the page" or "find the site" (both singular), while the informational topics indicated the plural, for example "find webpages".

Informational search topics

Informational search topics have varying numbers of relevant items located at different positions in the search results list. Table 7.1 shows the value of Average Precision for each topic. Table 7.3 shows the Click Precision, Click Recall and Click F-measure results, indicating how effectively users were able to identify relevant answers. The performance of each of the image and thumbnail interfaces was compared using an F-test. Click precision shows a statistically significant difference between the image and thumbnail interfaces ($F = 11.011, p = 0.0013$). In other words, for informational topics, users significantly manage to predict more relevant answers when using the image interface compared with the thumbnail interface.

The result of the Click Recall shows no significant difference in the number of relevant answers selected by users as a proportion of the total number of relevant answers available for that topic ($F = 2.0716, p = 0.1534$). However, the click F-measure, the harmonic mean between Click Precision and Click Recall, shows a statistically significant difference ($F = 5.4112, p = 0.0222$). The overall indication is that the salient image interface is effective and could often lead users to the correct answers for informational queries; in contrast, the thumbnail interface is less effective.

Navigational search topics

Navigational queries usually have only one possible answer page. In this study, 96 sessions involved navigational search topics where each subject answered two navigational topics using only one interface. Table 7.4 shows the total number of correct and incorrect choices made by users, split by interface.

Measures		Img	Thum
Click Precision	Average	0.8160	0.5590
	Stdev	0.3425	0.4129
	p-value	**0.0013**	
Click Recall	Average	0.6146	0.5069
	Stdev	0.3318	0.3979
	p-value	0.1534	
Click F-measure	Average	0.6667	0.5060
	Stdev	0.3053	0.3687
	p-value	**0.0222**	

Table 7.3: Click Precision, Click Recall and F-measure for user selection of the informational search topics.

Interface	Img	Thum	χ^2 p-value
Correct	27	44	**0.0436**
Incorrect	21	4	**0.0007**

Table 7.4: Total number of correct and incorrect answers selected by users for navigational search topics.

Analysis shows that users of the thumbnail interface were able to predict the correct answer with statistically significant frequency ($\chi^2, p = 0.0436$) and moreover, a statistically significant difference ($\chi^2, p = 0.0007$) was found in total number of the incorrect answers. This suggests that the thumbnail interface is effective and powerful for navigational search topics. This is in line with the findings of other researchers who considered re-finding tasks [Ayers and Stasko, 1995; Maarten et al., 1999; Dziadosz and Chandrasekar, 2002; Do and Ruddle, 2012].

7.2.2 Search completion time

In this study, task completion time is measured by three components: the time taken to select an answer; the total time required to answer the task; and the time required after selecting the answer before moving on to the next task. A two-way ANOVA is used to analyse the

total time taken based on the two categorical explanatory variables, interfaces (image and thumbnail) and type of search topic (informational and navigational).

Total time

The two-way ANOVA shows a statistically significant difference between the two types of search topics for the total time required to finish the task ($F = 3.9703, p = 0.0478$); but no significant difference was found between the two interfaces ($F = 1.6999, p = 0.1939$). These results show the significant impact of the topic type on task completion time, where users managed to finish answering the navigational search topics on average over 5 seconds faster than answering the informational topics. In contrast, the interfaces show no impact on the overall task completion time; this suggests that the interfaces did not aid users in answering the overall task types, but they may help in specific topic types (as outlined below).

Figure 7.2 shows the total time, in seconds, required for each combination of interface and topic types. Some outlier points occurred with the salient image interface; these users required extra time when browsing the result page. For the interaction effect (combinations of topic and interface), the two-way ANOVA shows a significant difference between combinations of interfaces and topic types for the total time required to finish answering the topics ($F = 8.3193, p = 0.0044$). Multiple comparisons were analysed using Tukey's HSD test; the results shown in Table 7.5 demonstrate that users managed to finish answering the navigational topics on average over 10 seconds faster using the thumbnail interface than with the image interface ($p = 0.0180$). In contrast, no statistically significant difference was found between the two interfaces for the informational topics ($p = 0.6791$). A statistical significant difference

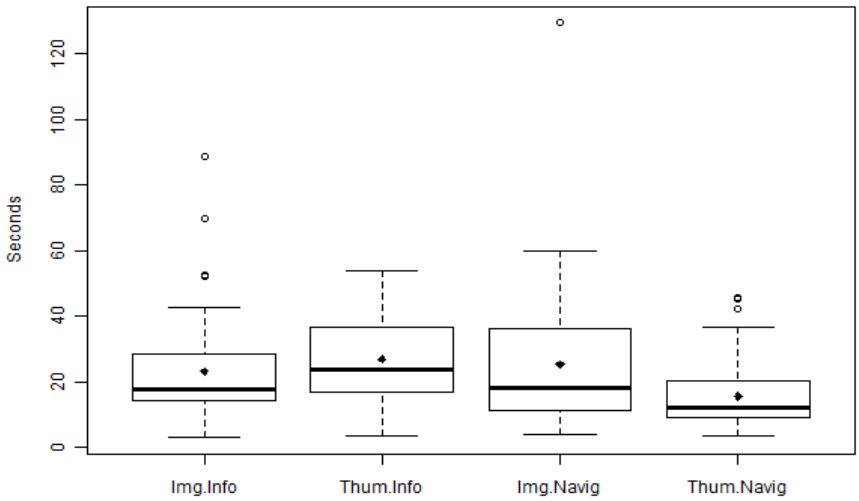

Figure 7.2: Total time required to answer the informational and navigational topics, split by interface.

Interface	Img:Info	Thum:Navig	Thum:Info
Thum:Navig	0.0947	-	-
Thum:Info	0.6791	**0.0039**	-
Img:Navig	0.9220	**0.0180**	0.9619

Table 7.5: Results of Tukey's HSD test for total time required to answer the search topics.

was found between the informational and navigational topics using thumbnail interface ($p = 0.0039$). (A possible reason might be the impact of topic type on thumbnail interface, we investigate this in Section 7.2.5 below.)

Time spent after selecting the answer

Time taken from selecting the answer to the end of the task shows how confident users were in judging the search results. Spending more time between selecting the answer and finishing the task may indicate that users were still viewing items and comparing selected

items with others. Results of a two-way ANOVA show no significant difference in the time spent after selecting the answer between the interfaces ($F = 0.9321, p = 0.3356$), but a significant difference was found between the two topic types ($F = 8.1369, p = 0.0048$). Thus, we may reasonably conclude that topic types have considerable impact on time spent after selecting answers, where user spent significantly longer time with informational topics than with navigational topics, see Figure 7.3.

One explanation for this significant difference is that with informational topics, more than one relevant answer might be presented for a topic, but with navigational topics there can be only one correct answer. Another possible explanation is that visual summaries provide better cues for navigational topics than for informational topics. However, results of a two-way ANOVA show no statistically significant difference in the interaction effect on the time required after selecting the answer before moving on to the next task ($F = 1.5386, p = 0.2164$). Similarly, in Section 6.2.2 of Chapter 6, results show no significant difference in the time spent after selecting the answer, when using navigational topics with thumbnail and image interfaces. This indicates that users with additional visual summaries are confident after selecting an initial answer.

7.2.3 User effort expended

Eye tracking enables us to capture the gaze position of users, which allow us to track the number of times the user has viewed a particular item whilst answering a given task. We collected the total number of uniquely viewed items and the percentage of items that were re-viewed.

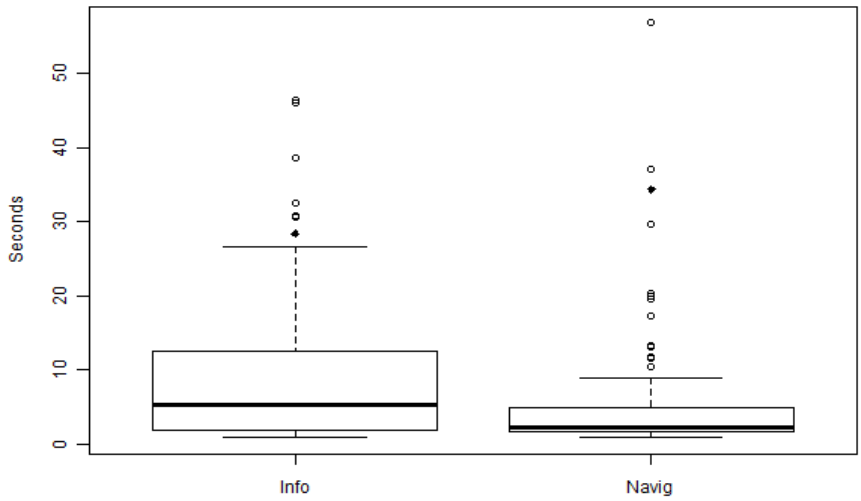

Figure 7.3: Total time required after selecting the answer before the end of tasks, split by topic types.

Uniquely viewed items: a search result was counted as viewed if the user viewed either the textual summary or visual summary, or both. The total number of uniquely viewed items is shown in Table 7.6, split by interface. Although the results show that users viewed more text for informational topics when using the thumbnail interface and, in contrast, users viewed more text for navigational topics when using the image interface, results showed no statistically significant differences for either the overall uniquely viewed items ($\chi^2, p = 0.9571$) or the uniquely viewed text summaries ($\chi^2, p = 0.1720$). As it can be observed that users view fewer text summaries for navigational topics than for informational topics, we statistically examined the difference in the textual unique viewed items for each interface. Results show a significant difference for the thumbnail interface only ($\chi^2, p = 0.0254$). This is an indication that for navigational topics, users are overall less dependent on viewing related

Search topic type	Items	Img	Thum
Informational	Overall uniquely viewed items (incl. visual summaries)	235	235
	Textual unique viewed items	205	225
Navigational	Overall uniquely viewed items (incl. visual summaries)	227	226
	Textual unique viewed items	204	180

Table 7.6: The total number of uniquely viewed items, split by interface.

text summaries.

Percentage of re-viewing: The total number of times that users viewed search result items, whether with textual or visual summaries or both, was collected to evaluate the percentage of re-viewing. The following formula was used to calculate the percentage of re-viewing of search result items for each session:

$$\text{Re-viewing} = \frac{Total\ viewed - Uniquely\ viewed}{Total\ viewed}$$

In other words, the formula measures the percentage difference between the amount total and unique views of search results. A two-way ANOVA shows no statistically significant difference between the two interfaces ($F = 0.1756, p = 0.6756$), between types of topics ($F = 0.9216, p = 0.3383$), or between interaction ($F = 2.3013, p = 0.1309$). These results suggest that the type of visual summaries and the topic types have no impact on the percentage of re-viewing search result items. This supports our previous findings in Chapter 6 (see Section 6.2.3) where results show no significant difference between image and thumbnail interfaces with navigational topics ($p = 0.9137$). The data suggests that users spent significantly less mental effort when additional visual summaries are presented.

7.2.4 User attention on specific informative components

To analyse user interaction with search results in more detail, we collected the attention spent on four informative components: visual summary, document title, snippet and URL. In this section, we evaluate the user's attention at the level of informative components to find the impact of the visual summary and topic type on the user attention paid to that particular component.

Visual attention distribution

It is essential to study the impact of the correlation between search topic types and existing visual summaries on user visual behaviour to understand in-depth user searching behaviour and the effectiveness of visual summaries. Visual attention is defined as the proportion of time spent viewing visual summaries out of the total time spent on text and visual summaries, where the gaze spent on regions of white-space between informative components is not included. Studying visual attention can therefore show the relationship between the search topic types (informational and navigational) and current approaches to visual summaries (image and thumbnail).

Figure 7.5 shows the percentage of visual attention spent on each interface, split by topic types. A two-way ANOVA is used to analyse the visual attention spent and the two categorical explanatory variables (interface and topic types). Results show no statistically significant differences between on overall visual attention spent on the two interfaces ($F = 2.8285, p = 0.0942$); see Figure 7.4. However, results show a statistically significant difference between topic types ($F = 4.1201, p = 0.0438$), where users spent a considerably larger amount

of time looking at visual summaries with navigational topics. Thus we may conclude that topic types have a significant impact on users' visual attention In addition, a statistically significant difference was found in the interaction between interfaces and topic types ($F = 5.0024, p = 0.0265$).

Consequently, pairwise comparisons were analysed using Tukey's HSD test as shown in Table 7.7. Figure 7.5 shows the distribution of visual attention spent on the interfaces, split by topic types. Results show that users spent a statistically significantly greater amount of attention on thumbnails with navigational topics than other combinations (thum:info ($p = 0.0153$), img:info ($p = 0.0459$) and Img:Navig ($p = 0.0310$)). This behaviour suggests that users found thumbnail summaries to be useful for navigational topics but not for informational topics.

In chapter 5, an F-test shows no significant difference between the four interfaces (thum, img, tag and VSnip) for informational topics. In contrast, results showed a significant difference on time spent on visual summaries between image and thumbnail interfaces with navigational topics in Chapter 6. One reason that may explain why the significant difference occurred in Chapter 6 but not in chapter 5 is the impact of topic types. Users spent significantly more time looking at visual summaries with navigational topics than with informational topics. The data thus shows that topic types have a strong impact on user visual attention, particularly in the thumbnail interface.

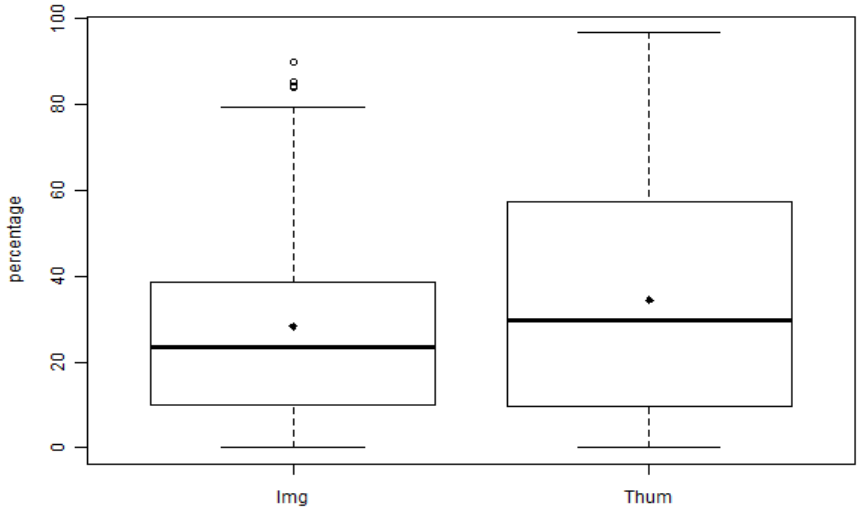

Figure 7.4: Percentage of visual attention spent on interfaces (Thum and Img).

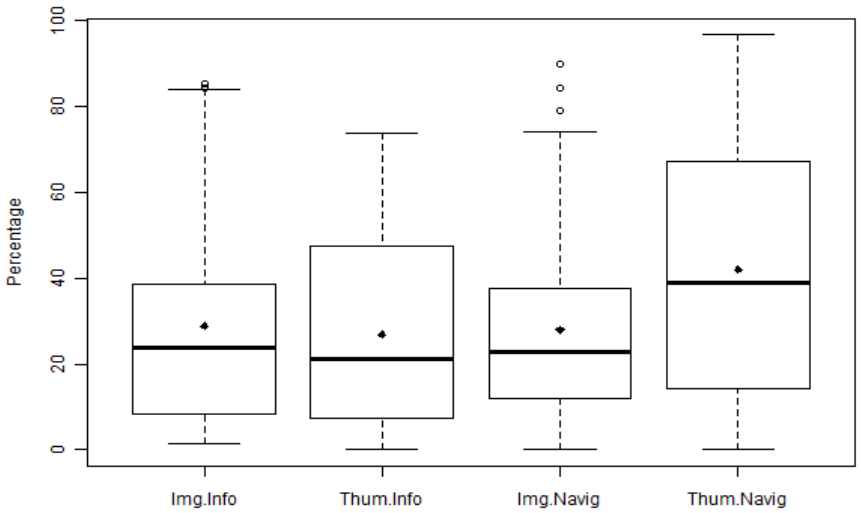

Figure 7.5: Percentage of visual attention spent, split by interface and topic type.

Interface	Img:Info	Thum:Navig	Thum:Info
Thum:Navig	**0.0459**	-	-
Thum:Info	0.9795	**0.0153**	-
Img:Navig	0.9989	**0.0310**	0.9947

Table 7.7: The results of Tukey's HSD test for visual attention, split by combination of interface and topic types.

Interaction with document title

The text summary for each item of the search results consists of three components: a document title, a URL, and a short text extract (snippet) from the source document. We collected the amount of time a user's gaze was focussed on each component of these text summary items, using the mask shown in Figure 7.1. We also identified the total time that users spent on each document title, and used a two-way ANOVA to analyse the data. The results show no significant differences between interfaces ($F = 2.6666, p = 0.1041$), nor between topic types ($F = 0.8193, p = 0.3665$), but a statistically significant interaction effect was present ($F = 0.8193, p = 0.0001$).

Figure 7.6 shows the total time spent viewing the text summary (document title), split by the combination of interface and topic types. Multiple comparisons were analysed using Tukey's HSD test to evaluate the combinations of interfaces and topic types – the results are shown in Table 7.8. Users spent a statistically significantly shorter time on document titles for navigational topics when thumbnails were presented, compared with the image interface for navigational topics ($p = 0.0006$), and the thumbnail interface for informational topics ($p = 0.0039$). In other words, users focus on the document title more when using the thumbnail interface, particularly for navigational topics.

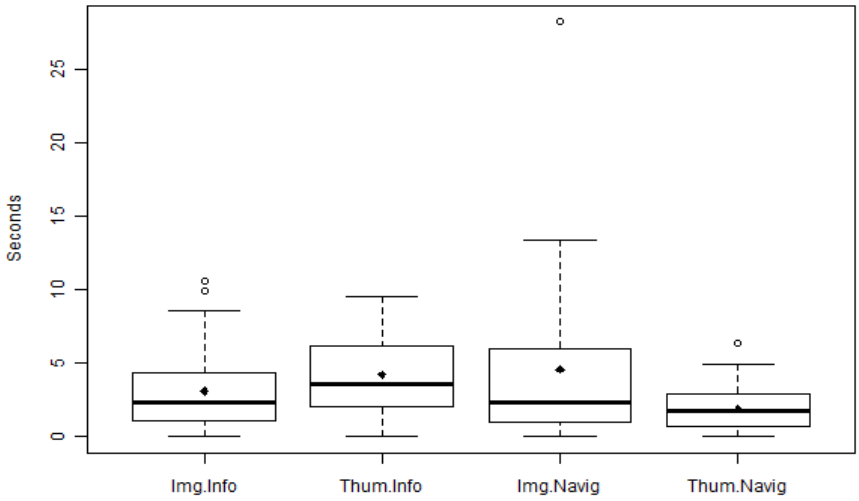

Figure 7.6: Total time spent on the textual component (document title), split by combination of interface and topic type.

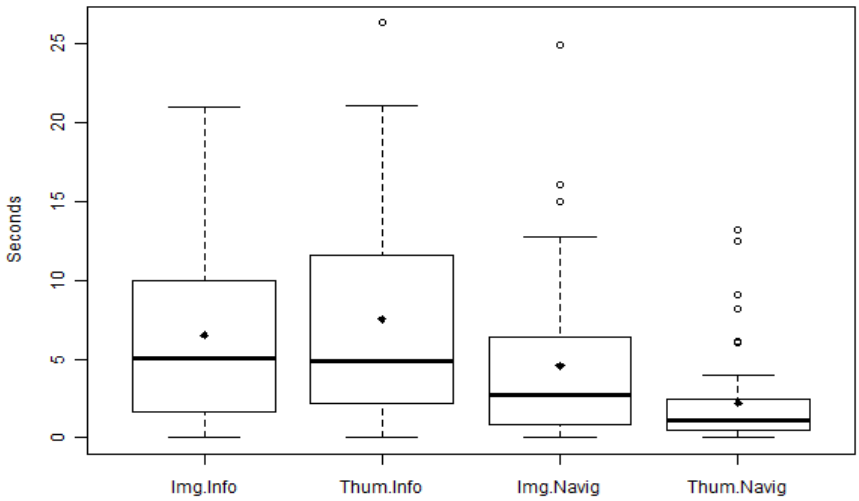

Figure 7.7: Total time spent on the textual component (short snippet), split by combination of interface and topic type.

Interface	Img:Info	Thum:Navig	Thum:Info
Thum:Navig	0.2791	-	-
Thum:Info	0.3515	**0.0039**	-
Img:Navig	0.1360	**0.0006**	0.9555

Table 7.8: The results of Tukey's HSD test for the total time spent on the textual component (document title), split by combination of interface and topic type.

Interaction with the snippet

We collected the duration of the user's gaze on each snippet for the two interfaces. Results of a two-way ANOVA show no significant difference between interfaces ($F = 0.8102, p = 0.36920$), but statistically significant differences were found between topic types ($F = 22.1787, p < 0.0001$) and interaction effect ($F = 4.9646, p = 0.0271$). Topic types significantly affect the gaze spent on snippets, where users spent a significantly larger amount of time looking at snippets with informational topics than with navigational topics.

For the interaction effect, Figure 7.7 shows the total time that users spent viewing the short text summary (snippet), split by the combination of interface and topic types.

Pairwise comparisons were analysed using Tukey's HSD test as shown in Table 7.9. The results confirm the significant impact of topic types, since users spent a significantly larger amount of time looking at snippets with informational topics in comparison with navigational topics, when using the thumbnail interface ($p < 0.0001$). The significant differences between the image interface with informational topics and the thumbnail interface with navigational topics ($p = 0.0005$) derive from the significant impact of topic types: this is the same for the significant difference between the thumbnail interface with informational topics and the image interface with navigational topics ($p = 0.0382$). However, a possible reason might be that additional visual summaries influence the duration of the user' gaze on snippets. We

Interface	Img:Info	Thum:Navig	Thum:Info
Thum:Navig	**0.0005**	-	-
Thum:Info	0.7839	$p < 0.0001$	-
Img:Navig	0.2987	0.1237	**0.0382**

Table 7.9: The results of Tukey's HSD test for the total time spent on the textual component (short snippet), split by combination of interface and topic type.

investigate this in more detail in Section 7.2.5 below.

Interaction with the URL

The length of time for which the users' gaze was fixated on the URL component of search result items was collected using the mask, as shown in Figure 7.1. A two-way ANOVA compares interfaces ($F = 0.0165, p = 0.8978$), topic types ($F = 12.4025, p = 0.0005$), and interaction effect ($F = 4.1656, p = 0.0427$). Results show that topic types significantly impact on the gaze duration spent on URL components, where users spent a significantly larger amount of time fixated on URL components with navigational topics than with informational topics. The total time spent viewing the URL components, split by the combination of interface and topic type, is shown in Figure 7.8. Multiple comparisons were analysed using Tukey's HSD test as shown in Table 7.10. Results show that for the image interface, users spent a significantly smaller amount of time on URLs with informational topics than with navigational topics ($p = 0.0006$). This can be explained by the fact that users are looking for a specific resource in navigational topics, and URLs show the domain website from which the result item was retrieved. The insignificant difference between informational and navigational topics with thumbnail interface ($p = 0.7219$) suggests that additional visual summaries may influence users' gaze distribution. We therefore investigate this in more detail in the next

Interface	Img:Info	Thum:Navig	Thum:Info
Thum:Navig	0.0807	-	-
Thum:Info	0.5309	0.7219	-
Img:Navig	**0.0006**	0.4192	0.0514

Table 7.10: The results of Tukey's HSD test for the total time spent on the textual component (URL), split by combination of interface and topic type.

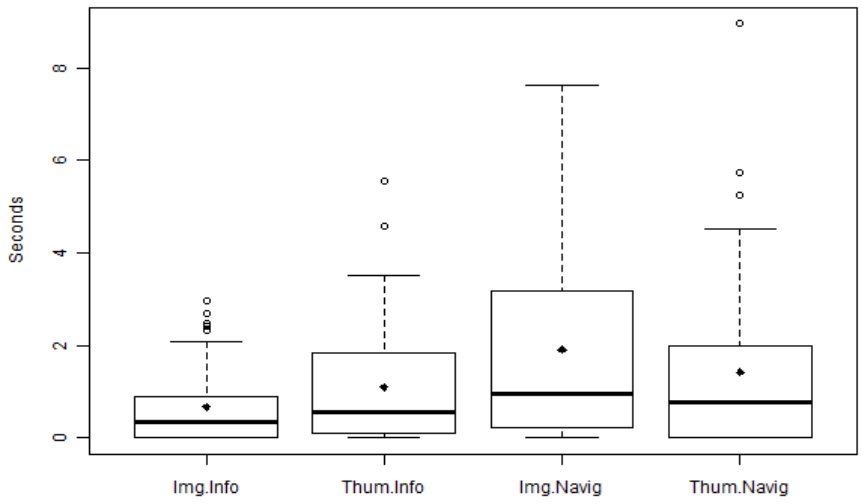

Figure 7.8: Total time spent on the textual component (URL), split by combination of interface and topic type.

section.

7.2.5 Comparison of user attention across different informative components

In the above sections, we compared the attention spent on a particular informative component to find the impact of visual summaries and topic types on attention paid to that particular component. In this section, we compare the gaze spent across the different informative components for the same interface, and for a particular topic type. Analysing

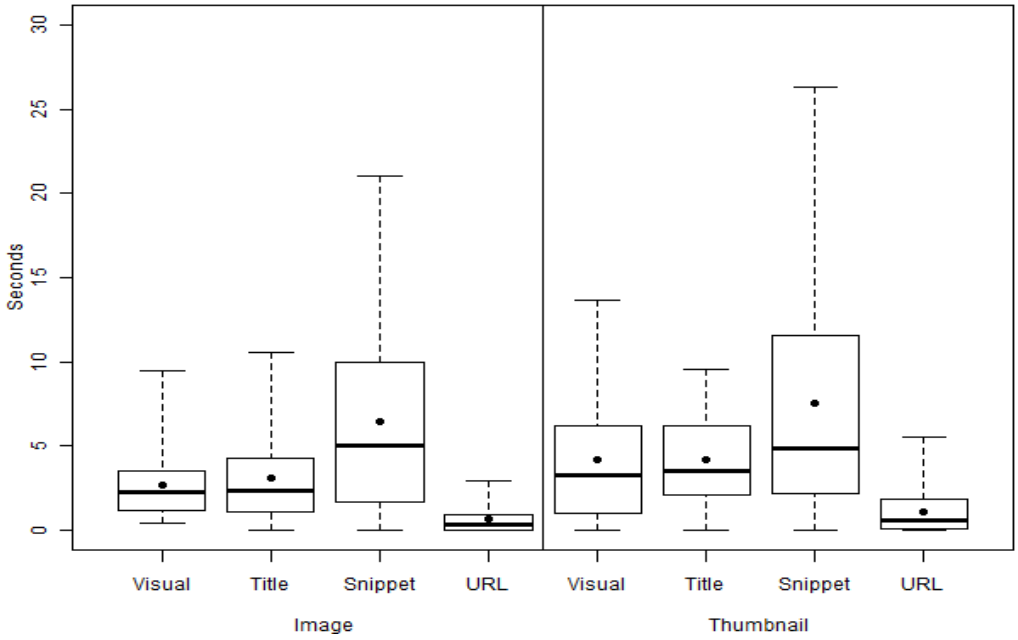

Figure 7.9: Total time spent on the four informative components of search result items, for informational topics on the thumbnail and image interfaces.

user attention across different components gives an in-depth understanding of the impact of visual summaries and topic types on the distribution of user attention between results screen components.

Impact of visual summaries on the distribution of users' gaze

In this section, we analyse the distribution of attention paid to the two interfaces for specific topic types separately, to investigate the impact of thumbnail and image summaries on user attention distribution.

Informational topics: Figure 7.9 shows the time spent on each one of the four informative components for the thumbnail and image interfaces, for informational topics. Table 7.11

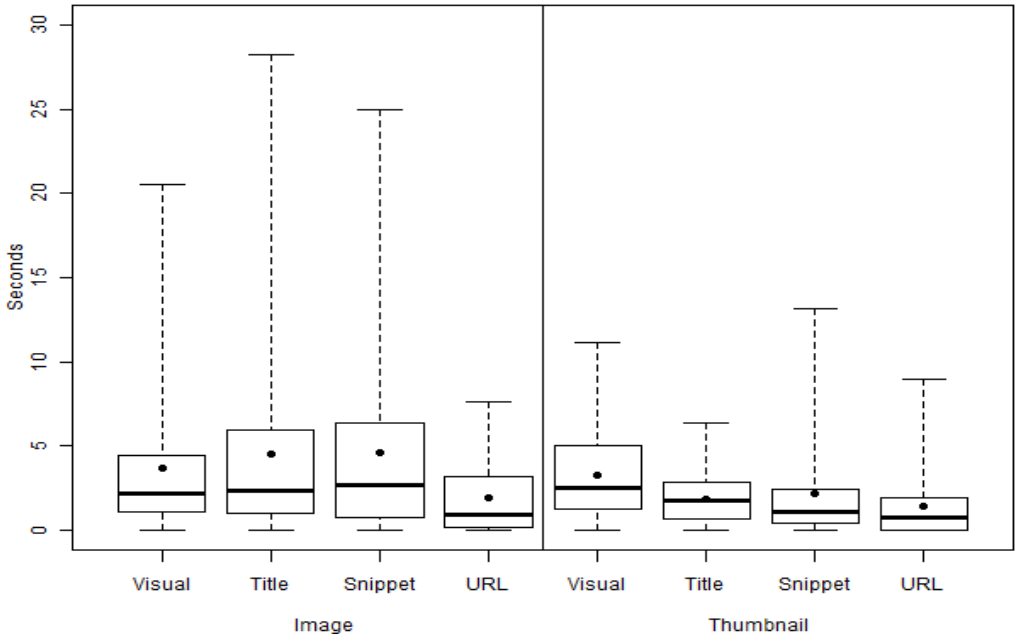

Figure 7.10: Total time spent on the four informative components of search result items, for navigational topics on the thumbnail and image interfaces.

shows the results of Tukey's HSD test for the total time spent on the four informative components, for informational topics. The results show similar trends for the two interfaces on the distribution of user gaze spent on the four informative components. For example, users spent significantly more time on text snippets on both interfaces, compared with other informative components ($p < 0.0001$). In contrast, no significant difference was found between the attention spent on document titles and visual summaries on both image ($p = 0.9468$) and thumbnail ($p = 1$) interfaces. This indicates that the type of visual summary did not influence user attention distribution among the four informative components.

Navigational topics: Figure 7.10 shows the time that was spent on each one of the four informative components for navigational topics, split by interfaces. Table 7.12 shows the

Interface	Components	Visual	Title	Snippet
Thum	Title	1	-	-
	Snippet	$p < 0.0001$	$p < 0.0001$	-
	URL	**0.0019**	**0.0020**	$p < 0.0001$
Img	Title	0.9468	-	-
	Snippet	$p < 0.0001$	$p < 0.0001$	-
	URL	**0.0134**	**0.0022**	$p < 0.0001$

Table 7.11: The results of Tukey's HSD test for the total time spent on the four informative components for informational topics.

results of Tukey's HSD test for the total time spent on the four informative components for navigational topics. Results show that visual summaries have a significant impact on the distribution of the user's gaze for navigational topics. For example, for the image interface, users spent significantly less time on URLs compared with document titles ($p = 0.0169$) or text snippets ($p = 0.0153$). In contrast, these differences were not significant for the thumbnail interface, when comparing the amount of time spent on URLs with time spent on document titles ($p = 0.7622$) and text snippets ($p = 0.3729$). Additionally, users spent significantly more time on thumbnail summaries compared with document titles ($p = 0.0182$) and URLs ($p = 0.0006$), but this difference was not significant for the image interface: document title ($p = 0.7383$) and URLs ($p = 0.2056$).

Impact of topic types on the distribution of users' gaze

In the previous section, we analysed the results in terms of the effect of visual summaries on the distribution of users' gaze. We now independently analyse the impact of topic types on each interface based on the results of Table 7.11 and Table 7.12, to understand the impact of topic types on the distribution of users' gaze at the level of informative components.

For the thumbnail interface, users spent a significantly larger amount of time on the

visual (thumbnail) component than on document titles for navigational topics ($p = 0.0182$); however this was not significant for informational topics ($p = 1$). In addition, users spent a significantly larger amount of time on text snippets than on document titles and URLs ($p < 0.0001$), but this was not significant for navigational topics.

For the image interface, users spent significantly more time on image summaries than on URLs for informational topics ($p = 0.0134$); in contrast, this was not significant with navigational topics ($p = 0.2056$). In addition, with the informational topics, users spent a significantly larger amount of time on text snippets than image summaries ($p < 0.0001$), but this was not significant for navigational topics ($p = 0.7197$).

The above results demonstrate that topic types have a significant impact on user gaze distribution at the component level. User attention was significantly different for specific informative components when particular topic types were compared with each other. With navigational topics, the distribution of users' gaze showed fewer significant differences than with informational topics. One reason for this difference is that users spent significantly more time looking at snippets than other informative components for informational topics. This suggests that users find that informative components are most useful with navigational topics, but for informational topics, users find snippets better than the other components.

7.3 Discussion and summary

In this chapter, we described a user study to evaluate the effectiveness of two visual summaries (thumbnail and image) and the impact of topic types on user searching behaviour, using twenty-four topics (12 informational and 12 navigational). Forty-eight subjects were asked

Interface	Components	Visual	Title	Snippet
Thum	Title	**0.0182**	-	-
	Snippet	0.0972	0.9201	-
	URL	**0.0006**	0.7622	0.3729
Img	Title	0.7383	-	-
	Snippet	0.7197	1	-
	URL	0.2056	**0.0169**	**0.0153**

Table 7.12: The results of Tukey's HSD test for the total time spent on the four informative components for navigational topics.

to answer a series of four topics (2 informational and 2 navigational) where each type of topic was answered using a different interface. This study differs from the studies described in Chapters 5 and 6, since in this chapter, we focus on analysing and comparing directly the impact of topic types on the effectiveness of additional visual summaries and user searching behaviour.

We used more topics (24 topics), and where topics were selected, various aspects (domains and length of the query string) and features (number and position of relevant answers) were taken into account, as shown in Tables 7.1 and 7.2. We also analysed the distribution of user gaze at the level of individual informative components, where a different mask was used to collect the user's gaze on the results screen.

We measured the completion time and the number of identified relevant and non-relevant selected answers for each task. For the informational topics, the image interface significantly improved the ability of users to predict relevant answers, but no significant difference was found between the two interfaces for task completion time. In Chapter 5, although the image interface required the least amount of time compared with other interfaces (text-only, thumbnail, visual tag and visual snippet), no significant difference was found between the

interfaces. One reason to explain the non-significant difference in Chapter 5 is that the number of topics (5 topics) was not enough to show the significance between the interfaces.

In contrast, for navigational topics, the thumbnail interface not only showed significantly better results in predicting relevant answers for navigational searches, but also helped users to finish navigational topics in a significantly shorter amount time than with the image interface. Our findings in Chapter 6 also show that the thumbnail interface significantly improved the performance of users, compared with other interfaces for navigational topics.

These results strongly indicate that topic types impact on the effectiveness of visual summaries. This suggests that search engines should show different visual summaries for different task types to make the user experience better.

Many studies have proposed approaches to classify query types [Kang and Kim, 2003; Kang, 2005; Beitzel et al., 2005]. They used different features to classify the query type; for example a phrase such as "where" refers to a navigational or transactional queries, while "what" precedes an informational query. A similar method could be used to select appropriate visual summaries to present on search results pages, based on the query type. Thus, salient images could be presented for informational queries, while the thumbnail is the best-performing visual summary for navigational topics. Google applied a similar technique to classify whether queries refer to a visual object or not, in order to present a new tool called "Knowledge Graph", see Section 2.4.4. Under their method, if the query is classified as a visual query, the "Knowledge Graph" is presented on the results page.

We also examined the unique views of text items and the percentage of re-viewing of search result items for both interfaces and topic types. The results showed that the two

interfaces have no impact on the amount of re-viewing of search result items. However, results show that users with navigational topics viewed significantly fewer text items than when using the thumbnail interface. This suggests that users of the thumbnail interface spent significantly less mental effort to find the answers for navigational topics, compared with the image interface.

In addition, we investigated the impact of topic types on user interaction with specific informative interface components (visual summaries, document titles, short text snippets and URLs). Results indicate that topic types impact significantly on user searching behaviour. Users spent more time looking at short text snippet components with informational topics compared with other informative components. In the navigational topics, user attention distribution was more balanced across the informative components. Results also show that topic types have a strong impact on visual summaries: for instance, users devoted significantly more attention to thumbnails with navigational topics than with informational topics. In the image interface, users gave more attention to image summaries when searching navigational topics than with informational topics. This suggests that visual summaries are more attractive to users with navigational queries than with informational queries. In contrast, results also suggest that users devote more attention to text summaries than visual summaries with informational queries.

These results provide good cues to present better search results. For informational queries, it would be helpful to provide more extensive text summaries to find relevant information, while for navigational queries more attention could be paid to improve visual summaries.

Chapter 8

Conclusion

The rapid increase in information provided on the World Wide Web makes it more complicated for users to select the right items among search results. Traditional search results focus mostly on textual summaries, yet newer visual techniques can also help. In this thesis, we present the findings of our in-depth investigation of the effects of visual summaries on user searching behaviour. Using eye tracking, we investigated user interactions with the informative components.

In the following sections, we summarize our contributions and discuss possible directions for future work.

8.1 Contributions

Visual summaries can be presented alongside text summaries, and can provide cues about the content of the retrieved web pages that may have positive impact on user searching behaviour and performance. Yet topic types may also influence user seeking behaviour and

the effectiveness of visual summaries. Eye tracking, as a tool, enables us to gain an in-depth understanding of how users interact with the results screen: thus we set out to address the following research questions:

- **Eye tracking.** How can an eye tracker be used to understand user behaviour when interacting with textual and visual summaries of search results?

  Using eye tracking in the evaluation of web search interfaces provided rich information on users' information search behaviour, particularly in the matter of user interaction with different informative components on a search results screen. In Chapter 3, we defined eye movement metrics that can be employed for the evaluation of web search interfaces. In addition, at the end of each user study, the eye tracking research question was revisited, and recommendations were summarised in Chapter 3. Building on these recommendations, some techniques were proposed to gain more information on user searching behaviour. For example, calculating the percentage of re-viewing (Section 3.4.5) provides detailed information about cognitive load (effort expended), whilst the use of a tracking diagram (Section 6.2.7) succinctly demonstrates the pattern of user interaction with the presented informative components of screen results.

  One of the main issues affecting the use of eye tracking in research is the quality of eye movements (calibration), as discussed in Section 3.5.2; therefore, we proposed a method in Section 3.7.3 that allows us to determine the quality of calibration, since the existing eye tracking system (Tobii Studio) does not provide any criteria for this aspect. Another issue is the adaptation of gaze direction, explained in Section 3.7.2. We used a black screen displaying for 3 seconds between screens to avoid the effect of

the previous screen on user gaze direction on the coming screen. A further issue when employing eye tracking in the evaluation of web search interfaces is the selection of the appropriate filter for the raw gaze-points data. In our studies, we filtered this data by removing noise data, identifying gaze points that occur in AOIs, optimising gaze data and identifying viewed AOIs, as described in Section 3.7.7.

- **Visual summaries.** Does providing additional visual summaries for the presentation of web search results impact on users' information-searching behaviour and performance?

  The investigation of this research question was conducted in two stages. In the first stage we evaluated visual representation in three publicly available search engines, where the interfaces vary primarily in the proportion of visual and text summaries displayed in search results. Our analysis indicates that most users spend a substantially larger proportion of time looking at text information than visual, and that those interfaces that focus on text-based representations of document content tend to lead to quicker task completion times for named-page finding search tasks. In contrast, when other specific task types are answered, results also show that search completion time varies greatly among interfaces, and an appropriate combination of textual and visual information leads to the shortest search completion time and the least number of wrong answers. Furthermore, the findings of this study provide us with a strong understanding of how to design a web search interface with appropriate features to run controlled experiments using eye tracking. For instance, elements such as dynamic features (for example auto-enlarging thumbnails when the mouse moves over them)

or scrolling pages distract the user's gaze, making it difficult to locate. To avoid this difficulty, such elements should not be employed.

In the second stage, based on the results of the first stage, five interfaces were designed: one text-only interface and four visual interfaces, where each visual interface presented a different approach to visual summaries (thumbnail, visual tag, salient image and visual snippet). This stage consisted of two studies, where we evaluated the effectiveness of the four types of visual summary with informational topics in the first study and navigational topics in a second study.

In the first study of the second stage, fifty participants carried out a series of searches for five informational topics using a different interface for each topic. The results show that visual summaries significantly impact on the behaviour of users, but not on their performance when predicting the relevance of answer resources. Users spend significantly less time looking at the textual components of summaries in the visual summary interfaces. Comparing users' ability to predict the relevance of answer pages with a text interface versus a visual interface suggests that the tested visual summaries can mislead users to select non-relevant items on informational search topics. However, the salient image interface provided better results for informational topics when compared with other visual interfaces.

In the second part of this investigation, another fifty participants carried out a series of five navigational topics using the different interfaces. We used different task types, aiming to find homepages and single web pages for navigational queries. Furthermore, we classified user behaviour based on visual attention, to better understand different

methods of browsing search results. The results show that apart from the salient image interface, users perform significantly better in terms of time and effort required to answer given search topics when additional visual summaries are presented. Our study also suggests that the more time the user spends on visual summaries, the greater the user's ability to correctly predict the relevance of answers. Less effort and time are then needed to find the required answer. Results also indicate that the thumbnail interface showed better results when compared with other visual summaries for navigational topics. In terms of results page browsing behaviour, different amounts of attention spent on looking at the additional visual summaries actually produce different forms of browsing.

- **Topic types.** How does the type of search topic influence the effectiveness of additional visual summaries for the presentation of web search results?

  Based on the findings of Chapters 5 and 6, we evaluate the best-performing visual interfaces: salient images for informational searches, and thumbnails for navigational searches. Twenty-four topics (12 navigational and 12 informational) were employed in this study to evaluate the impact of topic types on the effectiveness of thumbnail and image interfaces and the user's information seeking behaviour. Particularly, we evaluate the relationship between topic types and attention paid to different informative components for web search results.

  The results of this study confirm our previous findings where the salient image interface shows a better performance for informational topics, while thumbnails were more effec-

tive for the navigational topics. Users managed to finish navigational topics when using the thumbnail interface in a significantly shorter time than with the image interface. The differences in search success and task completion time between the different topic types across the two interfaces suggests that a significant correlation exists between the presented approaches of visual summary and topic types.

Results also show that topic types significantly impact on users' gaze distribution on the informative components for web search results. With informational topics, users spent significantly more time looking at text snippets compared with other informative components (visual summaries, document titles and URLs). In contrast, with navigational topics, users spent a significant amount of time on visual summaries when using the thumbnail interface, whilst on the image interface, no significant difference was found in the amount of time spent on text snippets compared with visual summaries and document titles. This suggests that topic types impact considerably on user seeking behaviour.

8.2 Future work

8.2.1 Eye tracking

Eye tracking has only recently been employed for the evaluation of web search interfaces. Therefore, proposing new techniques in this area has the potential to improve the process of evaluating web search interfaces.

Traditional IR metrics and eye movements

In web search interfaces, traditional information retrieval metrics such as Click Recall and Click Precision evaluate user success in finding desired information, but they cannot provide a clear view on user effort expended. In contrast, eye movement metrics, such as fixation duration and re-viewing rates, provide useful information on user interaction with the results screen and effort expended in searching. In other words, they evaluate user effort but cannot indicate the effectiveness of relevance prediction such as user search success with informational search tasks. Identifying the relation between those metrics (IR and eye movement metrics) and developing a combined model of searching behaviour would therefore provide improvements in using eye tracking in the evaluation of web search interfaces.

Cognitive processing

Information processing (cognitive process) consists of four operations – learning, problem solving, memory and comprehension – that assist an individual to make a decision to find a relevant answer for a web search task. An interesting future study would be to examine the correlation between the variety of individual differences in cognitive styles and user seeking behaviour, where equipment such as the eye tracker can provide data, particularly illuminating user attention spent on visual summaries.

8.2.2 Approaches to visual summaries

For the novel combination of a thumbnail and a tag cloud interface, the visual tag, text was used to allow users to get the gist of web page content before visiting the page. However,

it seems that the small font-size makes it difficult to read the presented visual summaries. An improvement can be made by enlarging the font size and limiting the number of words that are shown. Additionally, in this approach, text was generated based on simple of word frequency rates. Applying more advanced criteria such as those used in information extraction (IE) would enable the identification of further information from structured or unstructured documents, such as named entity recognition and terminology extraction. This could help users to recognise the content of the retrieved web page more easily than by viewing only the most frequent words.

8.2.3 Visual summaries and other topic types

The findings of this thesis show that each kind of visual summary can have some positive impact on user seeking behaviour for a specific topic type, but not for other topic types. Thus, more research could be conducted to select appropriate visual summaries to suit different topic types. In this thesis, we considered having task types for web search (classification based on user intents with a task), but there are many other possible approaches. One is to evaluate the effectiveness of different approaches to visual summaries with major categories based on users' search interests. For example, Spink et al. [2001] classified users' search interests into 11 categories, such as people, places or things, health or sciences and government. A future study could be conducted to evaluate the effectiveness of additional visual summaries on web search interfaces for these categories.

8.2.4 Impact of topic types on user searching behaviour

One contribution of this thesis is the finding that topic types impact significantly on the visual search behaviour of users, particularly on attention distribution amongst the informative components of the results screen. It would therefore seem useful to display different visual summaries depending on the type of search task. Some previous studies proposed methods for the classification of topic types [Kang and Kim, 2003; Kang, 2005; Beitzel et al., 2005]; then, based on these types, additional information can be provided for specific informative components. For instance, for informational topics, extra text can be provided for the snippet component, whilst for navigational topics, greater focus could be given to visual summaries, document titles and URLs.

8.3 Summary

In this thesis we have proposed some techniques to improve the use of eye tracking for the evaluation of web search interfaces. We have also investigated the impact of additional visual summaries on web search interfaces. Significant impacts on user seeking behaviour were found, in aspects such as search success, effort expended, and user strategies for browsing search results. We have also studied the impact of topic types on user searching behaviour and the effectiveness of additional visual summaries in web search interfaces. Topic types show impact considerably on the user's attention distribution across the presented informative components of the results screen. Topic types also influence the effectiveness of visual summaries, where users perform significantly better with navigational topics using the thumbnail interface, while with informational topics, the image interface performs significantly better.

The contributions made in this thesis help to provide a better understanding of the framework for using eye tracking in the evaluation of web search interfaces. In particular, the thesis considerably expands the existing understanding of how users interact with results screen components, and the impact of topic types when users are browsing those components. These contributions can help to improve the development of more effective web search interfaces, particularly for the presentation of visual summaries. These contributions can also help to optimise search results based on the improved understanding of the impact of topic types on users' search behaviour.

Appendix A

Glossary

A.1 Key measures

AOI	Area of interest.
Click Recall	The number of relevant answers selected by users as a proportion of the total number of relevant answers available for that topic (see Section 2.1.3).
Click Precision	The number of correctly identified relevant answers as a proportion of all answers that the user selected (see Section 2.1.3).
Click F-measure	Traditional information retrieval metric that calculates the harmonic mean between Click Precision and Click Recall (see Section 2.1.3).
Percentage of re-viewing	The percentage of times items were re-viewed by users (see Section 3.4.5).
Uniquely viewed items	A count of the number of search result items for which the user viewed either the accompanying textual summary or visual summary, or both (see Section 5.2.5).

A.2 Interfaces

Txt	Text-only interface that presents only text summaries consisting of a web page *title*, a *text snippet* (that is, a brief textual extract designed to relate the query terms to quotes from the source web page), and the *URL* of the underlying web page (see Section 5.1).
Thum	Thumbnail interface that displays the same summaries used in the text-only interface, but accompanied by a thumbnail, a screen shot of the retrieved web page, (see Section 5.1).
Tag	Visual tag interface that shows the same summaries used in the text-only interface, but accompanied by a visual tag, a novel combination of a thumbnail and a tag cloud, (see Section 5.1).
Img	Salient image interface that presents the same summaries used in the text-only interface, but accompanied by an excerpt image, a dominant image from an underlying document that is "relevant" to the user's query, (see Section 5.1).
VSnip	Visual snippet interface that presents the same summaries used in the text-only interface, but accompanied by a visual snippet, a combination of a logo and a salient image, (see Section 5.1).

Bibliography

A. Aaltonen, A. Hyrskykari, and K. Räihä. 101 spots, or how do users read menus? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 132–139. ACM Press/Addison-Wesley Publishing Co., 1998.

R. Abrams and J. Jonides. Programming saccadic eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 14(3):428–443, 1988.

R. Abrams, D. Meyer, and S. Kornblum. Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):529–543, 1989.

W. Albert. Do web users actually look at ads? A case study of banner ads and eye-tracking technology. In *Proceedings of the 11th Annual Conference of the Usability Professionals Association*, 2002.

H. Ali [Al Maqbali], F. Scholer, J. A. Thom, and M. Wu. User interaction with novel web search interfaces. In *Proceedings of the 21st Annual Conference of the Australian*

Computer-Human Interaction Special Interest Group: Design: Open 24/7, pages 301–304. ACM, 2009.

B. Allen. Cognitive differences in end user searching of a CD-ROM index. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–309. ACM, 1992.

P. Allopenna, J. Magnuson, and M. Tanenhaus. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of memory and language*, 38(4):419–439, 1998.

A. Aula, P. Majaranta, and K. Räihä. Eye-tracking reveals the personal styles for search result evaluation. pages 1058–1061. Springer, 2005.

A. Aula, R. Khan, Z. Guan, P. Fontes, and P. Hong. A comparison of visual and textual page previews in judging the helpfulness of web pages. In *Proceedings of the 19th international conference on World wide web*, pages 51–60. ACM, 2010.

E. Z. Ayers and J. T. Stasko. Using graphic history in browsing the world wide web. *World Wide Web*, pages 11–14, 1995.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*, volume 463. ACM press New York., 1999.

B. Bailey, C. Busbey, and S. Iqbal. Taprav: An interactive analysis tool for exploring workload aligned to models of task execution. *Interacting with Computers*, 19(3):314–329, 2007.

S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202, 2008.

S. M. Beitzel, E. C. Jensen, O. Frieder, D. Grossman, D. D. Lewis, A. Chowdhury, and A. Kolcz. Automatic web query classification using labeled and unlabeled training data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 581–582. ACM, 2005.

M. Betke, J. Gips, and P. Fleming. The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 10(1):1–10, 2002.

A. Bojko. Informative or misleading? heatmaps deconstructed. *Human-Computer Interaction. New Trends*, pages 30–39, 2009.

D. Bowman, R. Ortega, M. Hamrick, J. Spiegel, and T. Kohn. System and method for refining search queries, 2001. US Patent 6,169,986.

A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

D. Bruneau, M. Sasse, and J. McCarthy. The eyes never lie: The use of eye tracking data in hci research. In *Proceedings of the CHI*, volume 2, page 25, 2002.

M. P. Bryden and C. A. Rainey. Left-right differences in tachistoscopic recognition. *Journal of Experimental Psychology*, 66(6):568–571, 1963.

G. Buscher, E. Cutrell, and M. Morris. What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 21–30. ACM, 2009.

M. Byrne, J. Anderson, S. Douglass, and M. Matessa. Eye tracking the visual search of click-down menus. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*, pages 402–409. ACM, 1999.

M. Camilli, R. Nacchia, M. Terenzi, and F. Di Nocera. ASTEF: A simple tool for examining fixations. *Behavior research methods*, 40(2):373–382, 2008.

F. Campagnoni and K. Ehrlich. Information retrieval using a hypertext-based help system. *ACM Transactions on Information Systems (TOIS)*, 7(3):271–291, 1989.

C. Campbell and P. Maglio. A robust algorithm for reading detection. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–7. ACM, 2001.

R. Capra III and M. Pérez-Quiñones. Using web search engines to find and refind information. *Computer*, 38(10):36–42, 2005.

S. Card, J. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

B. L. Carol. Document representations and clues to document relevance. *Journal of the American Society for Information Science*, 49(14):1293–1303, 1998.

M. Castelhano and J. Henderson. The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, 34(3):660, 2008.

A. Chapanis. Theory and methods for analyzing errors in man-machine systems. *Annals of the New York Academy of Sciences*, 51(7):1179–1203, 1951.

H. Chen and S. Dumais. Bringing order to the web: Automatically categorizing search results. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 145–152. ACM, 2000.

A. Cockburn, S. Greenberg, B. McKenzie, M. Jasonsmith, and S. Kaasten. Webview: A graphical aid for revisiting web pages. In *Proceedings of the OZCHI*, volume 99, pages 15–22, 1999.

A. Cockburn, C. Gutwin, and J. Alexander. Faster document navigation with space-filling thumbnails. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 1–10. ACM, 2006.

K. Coffman and A. Odlyzko. The size and growth rate of the internet. *First Monday*, 3 (10-5), 1998.

A. Cohen. Car drivers' pattern of eye fixations on the road and in the laboratory. *Perceptual and motor skills*, 52(2):515–522, 1981.

M. Cole, J. Gwizdka, C. Liu, and N. Belkin. Dynamic assessment of information acquisition effort during interactive search. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011a.

M. Cole, J. Gwizdka, C. Liu, R. Bierig, N. Belkin, and X. Zhang. Task and user effects on reading patterns in information search. *Interacting with Computers*, 23(4):346–362, 2011b.

V. Coltheart. *Fleeting memories: Cognition of brief visual stimuli.* MIT Press, 1999.

L. Cowen, L. Ball, and J. Delin. An eye movement analysis of webpage usability. *People and Computers*, pages 317–336, 2002.

E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 407–416. ACM, 2007.

A. Divoli, M. Wooldridge, and M. Hearst. Full text and figure display improves bioscience literature search. *PLoS One*, 5(4):e9619, 2010.

S. Djamasbi, M. Siegel, and T. Tullis. Visual hierarchy and viewing behavior: an eye tracking study. *Human-Computer Interaction. Design and Development Approaches*, pages 331–340, 2011.

T. V. Do and R. A. Ruddle. The design of a visual history tool to help users refind information within a website. volume 7224 of *Lecture Notes in Computer Science*, pages 459–462. Springer Berlin Heidelberg, 2012.

H. Drewes. *Eye gaze tracking for human computer interaction.* PhD thesis, Ludwig-Maximilians-Universität München, 2010.

A. Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer, 2007.

S. Dziadosz and R. Chandrasekar. Do thumbnail previews help users make better relevance decisions about web search results? In *Proceedings of the 25th annual international ACM*

SIGIR conference on Research and development in information retrieval, pages 365–366. ACM, 2002.

N. Eger, L. Ball, R. Stevens, and J. Dodd. Cueing retrospective verbal reports in usability testing through eye-movement replay. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 129–137. British Computer Society, 2007.

C. Ehmke and S. Wilson. Identifying web usability problems from eye-tracking data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI... but not as we know it-Volume 1*, pages 119–128. British Computer Society, 2007.

R. Ekstrom, J. French, H. Harman, and D. Dermen. *Manual for kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service, 1976.

K. Ellis. Eye tracking metrics for workload estimation in flight deck operations. Master's thesis, 2009.

K. Ericsson and H. Simon. *Protocol analysis*. MIT press, 1985.

T. Falkmer, J. Dahlman, T. Dukic, A. Bjällmark, and M. Larsson. Fixation identification in centroid versus start-point modes using eye-tracking data. *Perceptual and motor skills*, 106(3):710–724, 2008.

R. Fidel. Qualitative methods in information retrieval research. *Library and Information Science Research*, 15:219–219, 1993.

B. Fischer and E. Ramsperger. Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, 57(1):191–195, 1984.

A. Flammer and W. Kintsch. Allocation of attention during reading, 1982.

D. Fontenot. Visual field differences in the recognition of verbal and nonverbal stimuli in man. *Journal of Comparative and Physiological Psychology*, 85(3):564–569, 1973.

L. Frazier and K. Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14 (2):178 – 210, 1982.

A. Friedman and L. Liebelt. On the time course of viewing pictures with a view towards remembering. *Eye movements: Cognition and visual perception*, pages 137–155, 1981.

F. Frische, J. Osterloh, and A. Lüdtke. Modelling and validating pilots visual attention allocation during the interaction with an advanced flight management system. *Human Modelling in Assisted Transportation*, pages 165–172, 2011.

J. Goldberg and J. Helfman. Comparing information graphics: a critical look at eye tracking. In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 71–78. ACM, 2010a.

J. Goldberg and J. Helfman. Scanpath clustering and aggregation. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 227–234. ACM, 2010b.

J. Goldberg and X. Kotval. Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6):631–645, 1999.

J. Goldberg, M. Stimson, M. Lewenstein, N. Scott, and A. Wichansky. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 51–58. ACM, 2002.

V. Goldberg. *The power of photography: How photographs changed our lives*. Abbeville, New York, 1991.

L. Granka and K. Rodden. Incorporating eyetracking into user studies at Google. In *Workshop Position paper presented at CHI*, 2006.

L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479. ACM, 2004.

P. Green. Where do drivers look while driving (and for how long). *Human factors in traffic safety*, pages 77–110, 2002.

H. Greene and K. Rayner. Eye movements and familiarity effects in visual search. *Vision Research*, 41(27):3763 – 3773, 2001.

K. Grill-Spector, T. Kushnir, T. Hendler, and R. Malach. The dynamics of object-selective activation correlate with recognition performance in humans. *Nature neuroscience*, 3:837–843, 2000.

Z. Guan and E. Cutrell. An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 417–420. ACM, 2007.

J. Gwizdka and I. Spence. What can searching behavior tell us about the difficulty of information tasks? a study of web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–22, 2006.

Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*, 2005.

R. Haber and M. Hershenson. *The psychology of visual perception.* Holt, Rinehart & Winston, 1973.

E. Hanson. Focus of attention and pilot error. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 60–60. ACM, 2004.

B. M. t. Hart, J. Vockeroth, F. Schumann, K. Bartl, E. Schneider, P. Knig, and W. Einhuser. Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition*, 17(6-7):1132–1158, 2009.

M. Hearst. Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 59–66. ACM Press/Addison-Wesley Publishing Co., 1995.

M. A. Hearst. *Search User Interfaces.* Cambridge University Press, 2009.

R. R. Hightower, L. T. Ring, J. I. Helfman, B. B. Bederson, and J. D. Hollan. Padprints: graphical multiscale web histories. In *Proceedings of the 11th annual ACM symposium on User interface software and technology*, UIST '98, pages 121–122, New York, NY, USA, 1998. ACM. ISBN 1-58113-034-1.

O. Hoeber and X. Yang. A comparative user study of web search interfaces: Hotmap, Concept Highlighter, and Google. pages 866–874, 2006.

L. Holm and T. Mäntylä. Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, 35(7):1664–1674, 2007.

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press, UK, England, 2011.

A. Hornof and A. Cavender. Eyedraw: enabling children with severe motor impairments to draw with their eyes. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 161–170. ACM, 2005.

A. Hornof and T. Halverson. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods*, 34(4):592–604, 2002.

W. Horrey and C. Wickens. In-vehicle glance duration: Distributions, tails, and model of crash risk. *Transportation Research Record: Journal of the Transportation Research Board*, 2018(-1):22–28, 2007.

Y. Huang and P. Gordon. Distinguishing the time course of lexical and discourse processes through context, coreference, and quantified expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):966, 2011.

P. Hughes and B. Cole. What attracts attention when driving? *Ergonomics*, 29(3):377–391, 1986.

J. Hyönä, R. Lorch Jr, and J. Kaakinen. Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology*, 94(1):44, 2002.

A. Inhoff and K. Rayner. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Attention, Perception, & Psychophysics*, 40(6):431–439, 1986.

T. Ishida and M. Ikeda. Temporal properties of information extraction in reading studied by a text-mask replacement technique. *Journal of the Optical Society of America A*, 6(10): 1624–1632, 1989.

R. Jacob and K. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 573–603, 2003.

B. Jansen and D. Booth. Classifying web queries by topic and user intent. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pages 4285–4290. ACM, 2010.

B. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: a study of user queries on the web. In *ACM SIGIR Forum*, volume 32, pages 5–17. ACM, 1998.

B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3): 1251–1266, 2008.

R. Jenkins, J. Beaver, and A. Calder. I thought you were looking at me direction-specific aftereffects in gaze perception. *Psychological Science*, 17(6):506–513, 2006.

F. Jennings, D. Benyon, and D. Murray. Adapting systems to differences between individuals. *Acta Psychologica*, 78(1):243–256, 1991.

N. Jhaveri and K. Räihä. The advantages of a cross-session web workspace. In *CHI'05 extended abstracts on human factors in computing systems*, pages 1949–1952. ACM, 2005.

B. Jiao, L. Yang, J. Xu, and F. Wu. Visual summarization of web pages. *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 499–506, 2010a.

B. Jiao, L. Yang, J. Xu, and F. Wu. Visual summarization of web pages. In *Proceedings of ACM SIGIR*, pages 499–506, 2010b.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM, 2005.

S. P. Johnson, D. Amso, and J. A. Slemmer. Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, 100(18):10568–10573, 2003.

H. Joho and J. M. Jose. A comparative study of the effectiveness of search result presentation

on the web. In *Proceedings of the 28th European conference on Advances in Information Retrieval*, ECIR'06, pages 302–313. Springer-Verlag, 2006.

H. Joho and J. M. Jose. Effectiveness of additional representations for the search result presentation on the web. *Information processing & management*, 44(1):226–241, 2008.

W. Jones, S. Dumais, and H. Bruce. Once found, what then? a study of keeping behaviors in the personal use of web information. *Proceedings of the American Society for Information Science and Technology*, 39(1):391–402, 2002.

S. Josephson and M. Holmes. Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 43–49. ACM, 2002.

M. Juhola, V. Jäntti, I. Pyykkö, M. Magnusson, L. Schalén, and M. Åkesson. Detection of saccadic eye movements using a non-recursive adaptive digital filter. *Computer methods and programs in biomedicine*, 21(2):81–88, 1985.

S. Kaasten, S. Greenberg, and C. Edwards. How people recognise previously seen web pages from titles, urls and thumbnails. *People and Computers*, pages 247–266, 2002.

Y. Kammerer and P. Gerjets. How the interface design influences users' spontaneous trustworthiness evaluations of web search results: comparing a list and a grid interface. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 299–306. ACM, 2010.

Y. Kammerer and P. Gerjets. Effects of search interface and internet-specific epistemic beliefs

on source evaluations during web search for medical information: an eye-tracking study. *Behaviour & Information Technology*, 31(1):83–97, 2012.

I.-H. Kang. Transactional query identification in web search. *Information Retrieval Technology*, pages 221–232, 2005.

I.-H. Kang and G. Kim. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71. ACM, 2003.

R. Karsh and F. Breitenbach. Looking at looking: The amorphous fixation measure. *Eye movements and psychological functions: International views*, pages 53–64, 1983.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3:1–224, 2009.

D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 74–75. ACM, 2002.

L. A. King. Visual navigation patterns and cognitive load. In *Proceedings of the 5th International Conference on Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: Held as Part of HCI International 2009*, FAC '09, pages 254–259. Springer-Verlag, 2009.

N. Kloth and S. Schweinberger. The temporal decay of eye gaze adaptation effects. *Journal of Vision*, 8(11):1–11, 2008.

O. V. Komogortsev, S. Jayarathna, D. H. Koh, and S. M. Gowda. Qualitative and quantitative scoring and evaluation of the eye movement classification algorithms. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10, pages 65–68. ACM, 2010.

E. Kowler. *Eye movements and their role in visual and cognitive processes*, volume 4. Elsevier Science Ltd, 1990.

R. Krishnapuram, H. Frigui, and O. Nasraoui. Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation. ii. *Fuzzy Systems, IEEE Transactions on*, 3(1):44–60, 1995.

R. Krovetz and W. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems (TOIS)*, 10(2):115–141, 1992.

H. Lam and P. Baudisch. Summary thumbnails: readable overviews for small screen web browsers. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 681–690. ACM, 2005.

G. Larsson. *Evaluation methodology of eye movement classification algorithms.* Skolan för datavetenskap och kommunikation, Kungliga Tekniska höskolan, 2010.

D. Lewandowski. Query types and search topics of german web search engine users. *Information Services and Use*, 26(4):261–269, 2006.

X. Li, Y.-Y. Wang, and A. Acero. Learning query intent from regularized click graphs.

In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 339–346. ACM, 2008a.

Z. Li, S. Shi, and L. Zhang. Improving relevance judgment of web search results with image excerpts. *Proceedings of the 17th international conference on World Wide Web*, pages 21–30, 2008b.

Y. Lin and W. Zhang. Evaluating interface usability based on eye movement and hand movement behavioral parameters. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 47, pages 653–657. SAGE Publications, 2003.

H. Liou. Assessing learner strategies using computers: New insights and limitations. *Computer Assisted Language Learning*, 13(1):65–78, 2000.

S. Littlefair, P. Brennan, W. Reed, M. Williams, and M. Pietrzyk. Does the thinking aloud condition affect the search for pulmonary nodules? In *SPIE Medical Imaging*. International Society for Optics and Photonics, 2012.

S. Liversedge and J. Findlay. Saccadic eye movements and cognition. *Trends in cognitive sciences*, 4(1):6–14, 2000.

F. Loumakis, S. Stumpf, and D. Grayson. This image smells good: effects of image information scent in search engine results pages. *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 475–484, 2011.

Maarten, M. P. Czerwinski, M. V. Dantzich, G. Robertson, and H. Hoffman. The contribution

of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3d. In *Human-Computer Interaction*, pages 163–170. Press, 1999.

N. Mackworth and A. Morandi. The gaze selects informative details within pictures. *Attention, Perception, & Psychophysics*, 2(11):547–552, 1967.

B. Manor and E. Gordon. Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *Journal of neuroscience methods*, 128(1):85–93, 2003.

H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Do users find looking at text more useful than visual representations? a comparison of three search result interfaces. *In Proceedings of the Australasian Document Computing Symposium*, pages 35–42, 2009.

H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Evaluating the effectiveness of visual summaries for web search. *The Fifteen Australasian Document Computing Symposium (ADCS2010), Melbourne, Australia*, pages 36–43, 2010.

H. A. Maqbali, F. Scholer, J. A. Thom, and M. Wu. Tracking the impact of visual summaries on navigational web search. *Manuscript submitted for publication to JASIST (Under revision to address reviewers' comments)*, 2012.

E. E. Marsh and M. D. White. A typographic analysis of turkish newspapers' websites. *Journal of Documentation*, 59(6):647–672, 2003.

M. Martens and M. Fox. Does road familiarity change eye fixations? a comparison between watching a video and real driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(1):33–47, 2007.

M. Masarakal. *Improving expertise-sensitive help systems.* PhD thesis, University of Saskatchewan, 2010.

M. E. Masson. Cognitive processes in skimming stories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5):400–417, 1982.

L. Maughan, S. Gutnikov, and R. Stevens. Like more, look more. look more, like more: The evidence from eye-tracking. *Journal of Brand Management*, 14(4):335–342, 2007.

R. Mayer, W. Bove, A. Bryman, R. Mars, and L. Tapangco. When less is more: Meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of educational psychology*, 88(1):64, 1996.

E. McQuarrie and B. Phillips. Indirect persuasion in advertising: How consumers process metaphors presented in pictures and words. *Journal of Advertising*, 34(2):7–20, 2005.

Q. Mei and K. Church. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 45–54. ACM, 2008.

A. L. Mendelson. For whom is a picture worth a thousand words: Effects of the visualizing cognitive style and attention on processing of news photos. *Journal of Visual Literacy*, 24: 1–22, 2004.

P. Messaris. Can pictures bridge cultures? In *Visual Persuasion The Role of Images in Advertising*, pages 90–128. SAGE Publications, 1996.

A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. Personalized search on the world wide web. *The Adaptive Web*, pages 195–230, 2007.

M. Mishkin and D. G. Forgays. Word recognition as a function of retinal locus. *Journal of Experimental Psychology*, 43(1):43–48, 1952.

N. Moacdieh and N. Sarter. Eye tracking metrics: A toolbox for assessing the effects of clutter on attention allocation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 56, pages 1366–1370. SAGE Publications, 2012.

K. K. Moe, J. M. Jensen, and B. Larsen. A qualitative look at eye-tracking for implicit relevance feedback. In *Proceedings of the Workshop on Context-Based Information Retrieval*, volume 326, pages 36–47, Roskilde, Denmark, 2007.

J. Nielsen. Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414. ACM, 1994.

J. Nielsen and K. Pernice. *Eyetracking web usability*. New Riders Pub, 2010.

D. Noton and L. Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision research*, 11(9):929–942, 1971.

M. Nyström and K. Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204, 2010.

K. Oertel and O. Hein. Identification of web usability problems and interaction patterns with the realeyes-ianalyzer. *Interactive Systems. Design, Specification, and Verification*, pages 311–319, 2003.

W. Ogden, M. Davis, and S. Rice. Document thumbnail visualizations for rapid relevance judgments: When do they pay off? In *The Seventh Text REtrieval Conference (TREC7)*, pages 528–534, 1998.

S. Outing and L. Ruel. The best of eyetrack iii: What we saw when we looked through their eyes. *Published on Poynter Institute (1/11/2012). Retrieved*, 20, 2004. URL www.poynterextra.org/eyetrack2004/main.htm/.

A. Palanica and R. Itier. Searching for a perceived gaze direction using eye tracking. *Journal of Vision*, 11(2), 2011.

B. Pan, H. Hembrooke, G. Gay, L. Granka, M. Feusner, and J. Newman. The determinants of web page viewing behavior: an eye-tracking study. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 147–154. ACM, 2004.

J. Peeck. Increasing picture effects in learning from illustrated text. *Learning and instruction*, 3(3):227–238, 1993.

H. Pekta. A typographic analysis of Turkish newspapers' websites. *Journal Academic Marketing Mysticism Online*, 5(2):271–280, 2012.

A. Pollatsek, K. Rayner, and J. Henderson. Role of spatial location in integration of pictorial information across saccades. *Journal of Experimental Psychology: Human Perception and Performance*, 16(1):199–210, 1990.

V. Ponsoda, D. Scott, and J. Findlay. A probability vector and transition matrix analysis of eye movements during visual search. *Acta psychologica*, 88(2):167–185, 1995.

A. Poole and L. Ball. *Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects*. IGI Global, 2005.

A. Poole, L. Ball, and P. Phillips. In search of salience: A response time and eye-movement analysis of bookmark recognition. *People and Computers XVIII Design for Life*, pages 363–378, 2005.

R. Radach. Definition and computation of oculomotor measures in the study of cognitive processes. *Eye guidance in reading and scene perception*, page 29, 1998.

R. Radach, J. Hyona, and H. Deubel. *The mind's eye: Cognitive and applied aspects of eye movement research*. North Holland, 2003.

R. Rao, G. Zelinsky, M. Hayhoe, and D. Ballard. Eye movements in iconic visual search. *Vision research*, 42(11):1447–1464, 2002.

K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.

K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8):1457–1506, 2009.

K. Rayner and M. Castelhano. Eye movements. *Scholarpedia*, 2(9):3649, 2007.

K. Rayner and A. Pollatsek. *The psychology of reading*. Lawrence Erlbaum, 1994.

K. Rayner, S. Liversedge, S. White, and D. Vergilino-Perez. Reading disappearing text cognitive control of eye movements. *Psychological science*, 14(4):385–388, 2003.

K. Rayner, S. Liversedge, and S. White. Eye movements when reading disappearing text: The importance of the word to the right of fixation. *Vision research*, 46(3):310–323, 2006.

E. Reichle, K. Rayner, and A. Pollatsek. The ez reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476, 2003.

E. D. Reichle, A. Pollatsek, D. L. Fisher, and K. Rayner. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125–157, 1998.

E. D. Reichle, A. Pollatsek, and K. Rayner. Ez reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, 7(1):4–22, 2006.

R. Reichle. Comparing the e-z reader model to other models of eye movement control in reading, 2000. URL http://cogprints.org/1169/.

R. S. Rele and A. T. Duchowski. Using eye tracking to evaluate alternative search results interfaces. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49 (15):1459–1463, 2005.

D. Richardson and M. Spivey. Eye tracking: Characteristics and methods. *Encyclopedia of Biomaterials and Biomedical Engineering.*, 2004.

L. Richstone, M. J. Schwartz, C. Seideman, J. Cadeddu, S. Marshall, and L. R. Kavoussi. Eye metrics as an objective assessment of surgical skill. *Annals of Surgery*, 252(1):177–182, 2010.

R. Riding and S. Rayner. *Cognitive styles and learning strategies: Understanding style differences in learning and behaviour*. D. Fulton Publishers, 1998.

D. Robinson. A method of measuring eye movemnent using a scieral search coil in a magnetic field. *IEEE Transactions on Bio-medical Electronics*, 10(4):137–145, 1963.

D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, pages 13–19. ACM, 2004.

D. Salvucci and J. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 142–151. Springer-Verlag New York, Inc., 1994.

M. Sanderson and C. Van Rijsbergen. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems (TOIS)*, 17(4):440–465, 1999.

D. Sauter, B. Martin, N. Di Renzo, and C. Vomscheid. Analysis of eye tracking movements using innovations generated by a kalman filter. *Medical and biological Engineering and Computing*, 29(1):63–69, 1991.

J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: an empirical

evaluation of clustered presentation approaches. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 2037–2040. ACM, 2009.

S. Schweinberger, N. Kloth, and R. Jenkins. Are you looking at me? neural correlates of gaze adaptation. *Neuroreport*, 18(7):693–696, 2007.

F. Seagull and N. Walker. The effects of hierarchical structure and visualization ability on computerized information retrieval. *International Journal of Human-Computer Interaction*, 4(4):369–385, 1992.

A. Sears and M. Young. Physical disabilities and computing technologies: an analysis of impairments. In *The human-computer interaction handbook*, pages 482–503. L. Erlbaum Associates Inc., 2002.

T. A. Shimoda. The effects of interesting examples and topic familiarity on text comprehension, attention, and reading speed. *The Journal of Experimental Education*, 61(2):93–103, 1993.

B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

B. Shneiderman. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 3–12. ACM, 2008.

B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools:

multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7. ACM, 2006.

S. Shrestha and K. Lenz. Eye gaze patterns while searching vs. browsing a website. *Usability News*, 9(1), 2007.

J. Sinclair and M. Cardew-Hall. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15–29, 2008.

A. Singhal. introducing the knowledge graph: things, not strings. *Official Google Blog, May*, 2012.

A. Smeaton. Progress in the application of natural language processing to information retrieval tasks. *The Computer Journal*, 35(3):268–278, 1992.

J. Smeets and I. Hooge. Nature of variability in saccades. *Journal of neurophysiology*, 90(1): 12–20, 2003.

R. Song, Z. Luo, J. Nie, Y. Yu, and H. Hon. Identification of ambiguous queries in web search. *Information Processing & Management*, 45(2):216–229, 2009.

A. Spink, D. Wolfram, M. B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3):226–234, 2001.

D. Sullivan. Webtop search rage study. *The Search Engine Report*, February 2001.

D. Sullivan. What is search engine spam? the video edition, 2008.

A. G. Sutcliffe, M. Ennis, and J. Hu. Evaluating the effectiveness of visual user interfaces for information retrieval. *International Journal of Human-Computer Studies*, 53(5):741–763, 2000.

J. Teevan, C. Alvarado, M. Ackerman, and D. Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 415–422. ACM, 2004.

J. Teevan, E. Adar, R. Jones, and M. Potts. Information re-retrieval: repeat queries in yahoo's logs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 151–158. ACM, 2007.

J. Teevan, S. Dumais, and D. Liebling. To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 163–170. ACM, 2008.

J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. Andr, and C. Hu. Visual snippets: summarizing web pages for search and revisitation. *Proceedings of the 27th international conference on Human factors in computing systems*, pages 2023–2032, 2009.

H. Terai, H. Saito, Y. Egusa, M. Takaku, M. Miwa, and N. Kando. Differences between informational and transactional tasks in information seeking on the web. In *Proceedings of the second international symposium on Information interaction in context*, pages 152–159. ACM, 2008.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10. ACM, 1998.

K. Vicente and R. Williges. Accommodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine Studies*, 29(6):647–668, 1988.

K. Vicente, B. Hayes, and R. Williges. Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(3):349–359, 1987.

E. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 171–180. ACM, 1993.

P. J. Waddill and M. A. McDaniel. Pictorial enhancement of text memory: Limitations imposed by picture type and comprehension skill. *Memory & Cognition*, 20(5):472–482, 1992.

C. Ware and H. Mikaelian. An evaluation of an eye tracker as a device for computer input 2. In *Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI/GI conference on Human factors in computing systems and graphics interface: Toronto, Ontario, Canada*, volume 5, pages 183–188, 1987.

M. Wedel and R. Pieters. Eye fixations on advertisements and memory for brands: A model and findings. *Marketing Science*, 19(4):297–312, Fall 2000.

P. Wetzel, G. Krueger-Anderson, C. Poprik, and P. Bascom. An eye tracking system for analysis of pilots' scan paths. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, volume 1996. NTSA, 1996.

R. Williams and R. Morris. Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology*, 16(1-2):312–339, 2004.

S. S. Won, J. Jin, and J. I. Hong. Contextual web history: using visual and contextual cues to improve web browser history. *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1457–1466, 2009.

A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli. Using thumbnails to search the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 198–205. ACM, 2001.

K. Wu, P. Yu, and A. Ballman. Speedtracer: A web usage mining and analysis tool. *IBM Systems Journal*, 37(1):89–105, 1998.

M. Wu, M. Fuller, and R. Wilkinson. Searcher performance in question answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 375–381. ACM, 2001.

S. Xu, H. Jiang, and F. C. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 83–90. ACM, 2008.

S. Xu, T. Jin, and F. Lau. A new visual search interface for web browsing. In *Proceedings of*

the second ACM international conference on web search and data mining, pages 152–161. ACM, 2009.

X.-B. Xue, Z.-H. Zhou, and Z. Zhang. Improve web search using image snippets. In *Proceedings of the 21st national conference on Artificial intelligence - Volume 2*, AAAI'06, pages 1431–1436. AAAI Press, 2006.

A. Yarbus, B. Haigh, and L. Rigss. *Eye movements and vision*, volume 2. Plenum press New York, 1967.

B. Yoo, J. Lea, and Y. Kim. The seamless browser: enhancing the speed of web browsing by zooming and preview thumbnails. In *Proceedings of the 17th international conference on World Wide Web*, pages 1019–1020. ACM, 2008.

L. Young and D. Sheena. Survey of eye movement recording methods. *Behavior Research Methods*, 7(5):397–429, 1975.

G. Zhu and G. Mishne. Mining rich session context to improve web search. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1037–1046. ACM, 2009.